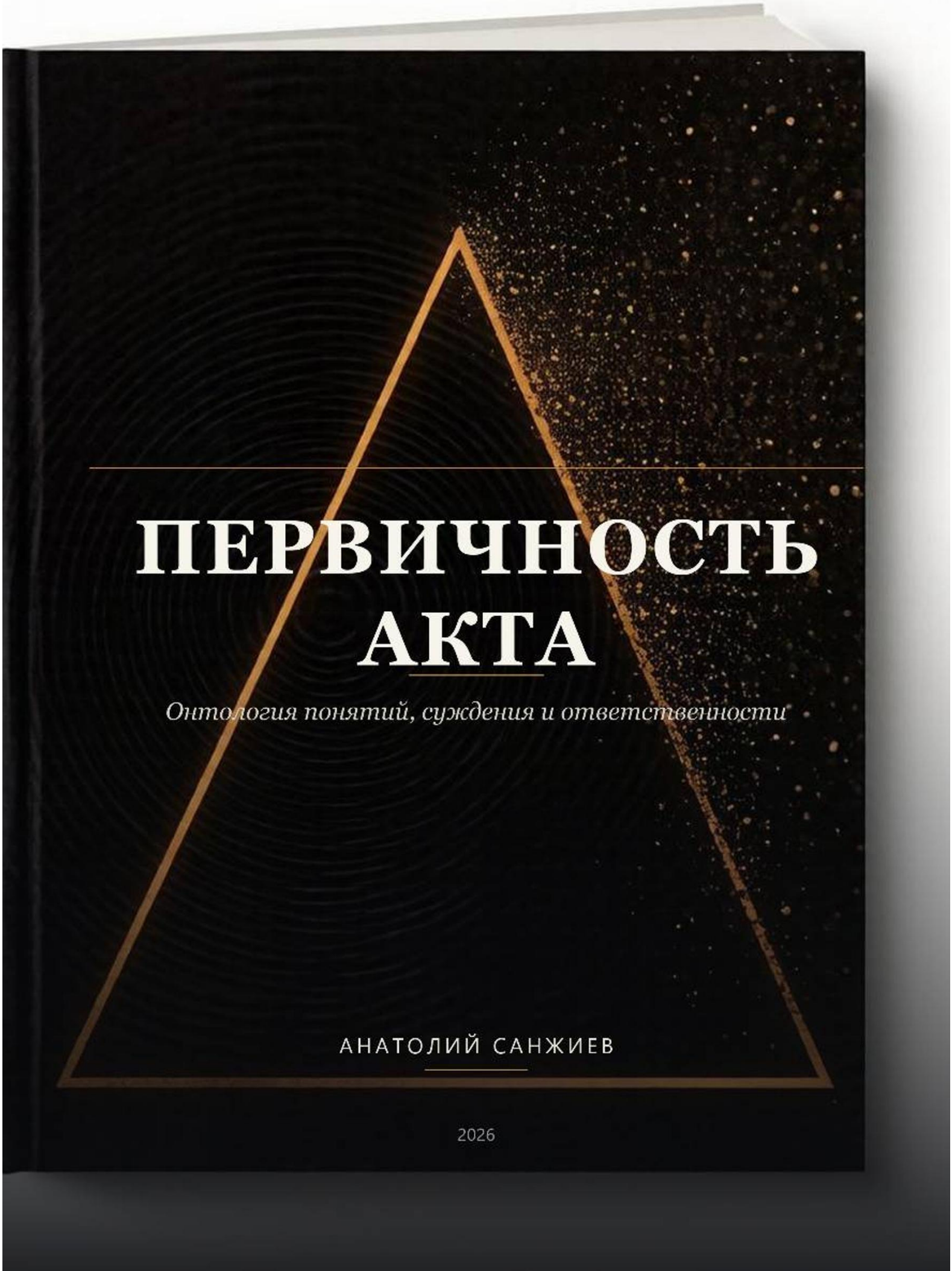

ПЕРВИЧНОСТЬ АКТА
Онтология понятий, суждения и ответственности
АНАТОЛИЙ САНЖИЕВ
2026

Анатолий Санжиев

Первичность акта

«Автор»
2026

Санжиев А. А.

Первичность акта / А. А. Санжиев — «Автор», 2026

Первичность акта: Онтология понятий, суждения и ответственностиЧто такое понятие? Мы привыкли, что это «вещь» в голове. Автор книги предлагает радикальный сдвиг: понятие — это не то, что мы имеем, а то, что мы делаем.Опираясь на традицию от Аристотеля до Канта и Выготского, автор строит триадическую модель акта: селекция, идентификация, проекция и обязательство. Понятие — не ментальная мебель, а стабилизировавшаяся компетенция, связывающая нас с миром.Из этого следуют выводы для эпистемологии (истина как инвариантность действия), этики (физика антиципации), права (функциональная ответственность машин) и нейронаук (модель «сломанного двигателя» при постинсультной апатии).Книга для философов, юристов, инженеров и клиницистов. Это не обзор, а позиция. Автор честно указывает, что осталось недоказанным, и формулирует условия, при которых теория была бы опровергнута.Если вы задумывались, почему «знаю, но не делаю» — не лень, а структурная проблема, — эта книга для вас.

© Санжиев А. А., 2026
© Автор, 2026

Содержание

ПЕРВИЧНОСТЬ АКТА 7
Предисловие 8
Как читать эту книгу 10
Глава 1. Вопрос, который так и не был задан 11
Акт, действие, событие 14
Цена операционального вопроса 16
Глава 2. Зеркало и его критики 17
Деятельностная традиция 20
Кантианское наследство и его долг 22
Глава 3. Модели, отчёты и иллюзия прозрачности 24
Идеализация без модели 27
Глава 4. Идеальные репрезентационисты 28
Конец ознакомительного фрагмента. 30

Анатолий Санжиев
Первичность акта

АНАТОЛИЙ САНЖИЕВ
ПЕРВИЧНОСТЬ АКТА
Онтология понятий, суждения и ответственности
Екатеринбург · 2026
Научное издание

Что если единицей познания является не представление, а акт? Книга защищает одну мысль: агент выделяет область ситуации, определяет её аффордансы, проецирует последствия и связывает себя с ними обязательством. Из этой структуры выводятся объективность как инвариантность, этика как физика антиципации, функциональная ответственность и архитектурные требования к искусственным системам.

© Анатолий Санжиев, 2026

ISBN присваивается Ridero

Все права защищены. Никакая часть книги не может быть воспроизведена без разрешения правообладателя.

Содержание

ПЕРВИЧНОСТЬ АКТА 6
Предисловие 7
Глава 1. Вопрос, который так и не был задан 12
Глава 2. Зеркало и его критики 22
Глава 3. Модели, отчёты и иллюзия прозрачности 32
Глава 4. Идеальные репрезентационисты 39
Глава 5. Селекция: цена внимания 46
Глава 6. Идентификация: аффордансы вместо свойств 52
Глава 7. Проекция: антиципация без симуляции 58
Глава 8. Обязательство: акт, который связывает 68
Глава 9. Ошибка, диссонанс и рефлексивный поворот 75
Глава 10. Объективность без объектов 82
Глава 11. Этика как физика антиципации 90
Глава 12. Конфликт актов 98
Глава 13. Ответственность без лиц 105
Глава 14. Институты, сети и время 114
Глава 15. Искусственный акт 121
Глава 16. Инженерия триады 127
Глава 17. Квантовый соблазн 137
Глава 18. Что осталось недоказанным 142
Заключение. Акт после акта 150
Приложение A. Глоссарий 156
Приложение B. Формальное ядро 160
Приложение C. Библиография 162

Перечень иллюстраций

Рис. 1. Две картины познания: репрезентационистский разрыв и акт как само отношение.
Рис. 2. Триада и её связывание: селекция, идентификация, проекция, обязательство.
Рис. 3. Акт первого порядка и следующая за диссонансом реконструкция второго порядка.

Рис. 4. Инвариантность при расширении практики: локальная привычка ломается, компетенция выдерживает.

Рис. 5. Этический диссонанс как разрыв между предполагаемой и действительной целостностью.

Рис. 6. Пять критериев функциональной ответственности.

Рис. 7. Петля Triadic-AC: акт, замыкаемый последствиями, а не токенами.

ПЕРВИЧНОСТЬ АКТА

Онтология понятий, суждения и ответственности

Анатолий Санжиев

Предисловие

У этой книги один тезис и множество следствий. Тезис состоит в том, что первичной единицей познания, суждения и ответственности является акт, а не репрезентация. Всё остальное на этих страницах — попытка увидеть, насколько далеко можно провести это единственное утверждение, не смошенничав, то есть не возвращая тихо и незаметно картину ума как вместилища внутренних картинок.

Аргумент начинается не с определения акта. Он начинается с диагноза. Около двадцати четырёх столетий центральный вопрос эпистемологии формулировался как вопрос о предметах: что такое понятие, что такое пропозиция, что такое ментальное состояние и каким образом оно ухитряется быть о чём-то другом. Я буду доказывать, что вопрос поставлен неправильно — не потому, что ответы были неверны, а потому, что сама форма вопроса предопределяет целое семейство ответов, в которых понятие, суждение и убеждение предстают как вещи, которыми агент обладает и которые затем применяет. Как только понятие становится вещью, которой обладают, отношение между этой вещью и миром превращается в неразрешимую головоломку — в разрыв, который затем заполняют причинными цепями, конвенциями, причинно-историческими цепями референции или простым декретом. Каждый такой мост — это репрезентация под новым именем.

Альтернатива — не новый мост, а другая отправная точка. Вместо того чтобы спрашивать, что такое понятие, спросим, что делает агент, когда понятие становится действующим. Ответ, который я защищаю, имеет структурный характер: агент выделяет область своей ситуации, идентифицирует то, что эта область ему предоставляет, проецирует траекторию последствий и тем самым связывает себя с миром, который теперь определён для него определённым образом. Селекция, идентификация, проекция, обязательство. Эти четыре фазы не являются этапами последовательности, которую можно было бы завершить где-то в другом месте; они суть внутренняя расчленённость единого временного акта. Понятие, с этой точки зрения, есть стабилизировавшийся акт — компетенция, ставшая достаточно надёжной, чтобы её можно было назвать, но остающаяся актом в самом способе своего существования.

Отсюда следуют четыре вывода, и они организуют четыре части книги.

Во-первых, эпистемологию нужно перестраивать вокруг успеха и неудачи, а не вокруг соответствия. Если знать — значит определённым образом действовать, то критерием знания служит инвариантность акта при расширении практики: одна и та же компетенция должна продолжать работать, когда ситуация разрастается, когда появляются новые случаи, когда меняются инструменты. Это критерий более строгий и более честный, чем соответствие, потому что он держит ответ перед тем, что действительно происходит с агентами, и не требует сравнения с божественной точки зрения между репрезентацией и миром, который эта репрезентация якобы представляет.

Во-вторых, суждение есть не приписывание свойства, а конституирование ситуации. Судить, что это вода, — значит не наклеить ярлык на образец, а завершить исследование, обязавшись миром, в котором целый круг антиципаций теперь получает основание, а другие исключаются. Вот почему суждение есть нормативный акт, а не только описательный, и вот почему пресловутый разрыв между сущим и должным уже, чем кажется: акт суждения уже содержит в себе направление обязательства.

В-третьих, ответственность архитектурна прежде, чем моральна. Агент ответствен, в функциональном смысле, разрабатываемом здесь, когда его архитектура позволяет ему регистрировать причинную связь между собственным актом селекции и тем исходом, который он предвидел, а также регистрировать диссонанс, когда исход расходится с предвиденным. Это проверяемо, распределяемо и не зависит от того, испытывает ли кто-нибудь чувство вины. Это

единственное понятие ответственности, которое можно применить к искусственным системам без фикции и без уклонения.

В-четвёртых, вопрос о машинах — это вопрос не о сознании, а об акте. Система, которая совершает триадический акт и связана своими исходами, есть субъект в единственном смысле, который производит онтологическую работу: она участник конституирования общего мира, а не инструмент, случайно производящий предложения.

Книга написана для философов и занимает позиции. Она не обозревает литературу ради удовольствия от обозрения. Там, где традиция права, я это говорю и строю дальше на ней; там, где она ошибается, я объясняю почему и стараюсь сформулировать возражение как можно сильнее, прежде чем на него ответить. Главы 1–4 расчищают почву. Главы 5–9 строят триаду и её внутреннее устройство — включая то, что большинство теорий действия упускает, а именно момент обязательства. Главы 10–14 распространяют этот подход на объективность, этику, право, институты и время. Главы 15–18 обращаются к машинам, к инженерии, к искушениям физики и, наконец, к вопросу, на который обязан ответить любой честный строитель систем: что показало бы, что всё это неверно.

Два предупреждения о том, чем эта книга не является. Это не теория сознания. Акт, как я его описываю, может совершаться без феноменального опыта, и я не утверждаю, что его совершение порождает опыт. Это также не законченная формальная теория. Центральное количественное понятие — этический диссонанс — определено здесь лишь с точностью до монотонности, и в главе 18 я прямо говорю, что метрика пока не построена. Этот пробел не риторическая уступка; это самая важная открытая проблема программы, и делать вид, что это не так, значило бы испортить весь аргумент.

Несколько слов об отношении этой книги к серии статей, на которые она опирается. Серия писалась как вереница вмешательств, каждое из которых было адресовано определённой аудитории — эпистемологам, логикам, теоретикам права, инженерам, клиницистам. Книга так работать не может. То, что следует далее, — единый аргумент в восемнадцати главах, с техническим аппаратом, вынесенным в приложение, и с полемикой, подчинённой конструкции. Там, где статьи и книга расходятся, книга точнее указывает на то, что я думаю на самом деле.

Как читать эту книгу

Книга — один аргумент в четырёх частях, и входить в неё можно с разных сторон. Кто хочет целиком — читает по порядку. У кого есть свой интерес — идёт своим маршрутом.

Философам — части I и II вплотную: там онтологическое утверждение формулируется и защищается. Часть III покажется приложением, и это не так: в ней критерий части II проверяется на самых трудных случаях — конфликт, ответственность, институты. Часть IV можно читать выборочно: глава 15 следует из онтологии, глава 16 — инженерия, глава 17 — отрицательный пример, глава 18 — аудит. Философ, который хочет сразу возражение, пусть начнёт с главы 18: книга устроена так, чтобы выдержать и такой порядок.

Юристам и регуляторам — начинать с главы 13, затем главы 8 и 12, и только потом вернуться к главам 5–7 за словарём. Правовой аргумент не предполагает всей онтологии, но предполагает понятие обязательства, а оно строится в главе 8.

Тем, кто занимается ИИ и инженерией, — начинать с главы 4, где диагноз дан на языке области, затем главы 7, 15 и 16. Обозначения собраны в приложении. Философские главы не предварительная часть, которую можно пропустить: архитектура главы 16 непонятна без структуры фаз, а структура фаз — это и есть аргумент части II.

Три соглашения. Первое: слова «акт», «действие» и «событие» употребляются в разных смыслах, разъяснённых в разделе «Акт, действие, событие» главы 1. Второе: формальный аппарат намеренно тонок — единственные математические утверждения суть упорядочения и монотонность, а формулы собраны в приложении B, чтобы прозу можно было читать без них. Третье: книга не обозревает литературу. Где традиция уместна, она включена в разговор; где нет — оставлена в покое.

Книга заканчивается там, где должна заканчиваться философия: условиями, при которых она была бы опровергнута.

Глава 1. Вопрос, который так и не был задан

История теории понятий есть история превосходных ответов на неправильный вопрос. От Аристотеля до теории прототипов вопрос оставался одним и тем же: что такое понятие? Абстрактный ли это предмет, ментальная единичность, правило, прототип, набор выводных способностей, узел в семантической сети? Каждый ответ защищается доводами значительной изощрённости, и каждый из них в конце концов наталкивается на один и тот же класс трудностей — трудность объяснить, каким образом определённая внутренняя или абстрактная единица ухитряется быть о некоторой определённой части мира и каким образом она применяется к случаю, которого сама в себе не содержит.

Картина провала на удивление устойчива. Предположим, понятие есть набор необходимых и достаточных условий. Тогда применение понятия к новому случаю требует дополнительной способности: распознать, что случай удовлетворяет этим условиям. Если эта способность сама понятийна, мы получаем регресс; если нет — мы вводим необъяснённую непонятийную способность. Предположим вместо этого, что понятия суть прототипы, градуированные структуры с центром и периферией. Тогда та же проблема возвращается в другом месте: агент должен распознать сходство между случаем и прототипом, а сходство не есть свойство случая самого по себе, но отношение, зависящее от того, что агент делает. Предположим, наконец, что понятия суть выводные роли, определяемые их местом в структуре обязательств и полномочий. Тогда вопрос превращается в вопрос о том, каким образом структура выводных ходов закреплена в мире, и ответ — что некоторые ходы ограничены восприятием — вводит обратно само понятие невыводного данного, критику которого инференциалист начинал с того, что провозгласил своей отправной точкой.

Это не череда случайностей. Это симптом. Вопрос «что такое понятие?» приглашает к ответу в категории предмета: понятие есть нечто существующее, чем обладают и что применяют. Как только эта грамматика установлена, отношение между понятием и миром должно быть отношением между двумя элементами, а всякое такое отношение либо причинно (и тогда слишком слабо, чтобы определять содержание), либо конвенционально (и тогда слишком произвольно, чтобы определять истину), либо магично (и тогда вовсе не теория).

Альтернатива, которую я предлагаю, — не новый ответ в той же категории. Это смена вопроса. Вместо того чтобы спрашивать, что такое понятие, спросим, что делает агент, когда понятие становится действующим: что должно быть верно об агенте, ситуации и исходе, чтобы фраза «у неё есть понятие воды» была верна о ней в определённый момент времени. Сдвиг этот грамматический, но его последствия онтологичны. Если обладать понятием — значит нечто делать, то понятие не есть единица, опосредующая отношение агента и мира; оно есть способ, которым организована деятельность агента. А если оно есть способ действовать, то оно имеет структуру действия: оно временно, оно укоренено в ситуации, оно исправимо и оно завершается обязательством.

Я называю это операциональным вопросом. У него три преимущества перед вопросом о предмете. Во-первых, на него можно ответить, наблюдая за исполнением, в том числе за исполнением систем, которые не являются ни сознающими, ни человеческими; это важно в части IV, но важно уже здесь, потому что теория понятий не должна молча ограничиваться существами с определённого рода внутренней жизнью. Во-вторых, он делает ошибку центральной, а не периферийной: если понятие есть акт, то заблуждение есть неудавшийся акт, и анализ неудачи даёт структурную информацию о том, чего требует успех. В-третьих, он объясняет, почему понятия поддаются нормативной оценке без обращения к особому царству значений: акт может быть компетентным или некомпетентным, а мерила компетенции внутренни для практики, в которой акт совершается.

Цена этого сдвига в том, что мы лишаемся утешения определения, умещающегося в одно предложение. Мы не можем сказать: «понятие есть X», где X — вид сущности. Мы можем сказать лишь, чем понятие является как акт: стабилизировавшейся компетенцией структурировать ситуацию так, чтобы это поддерживало антиципацию. Это менее аккуратно и в то же время более информативно, потому что говорит нам, чего искать, когда мы хотим знать, наличествует ли понятие, — и чего искать, когда мы подозреваем, что его нет.

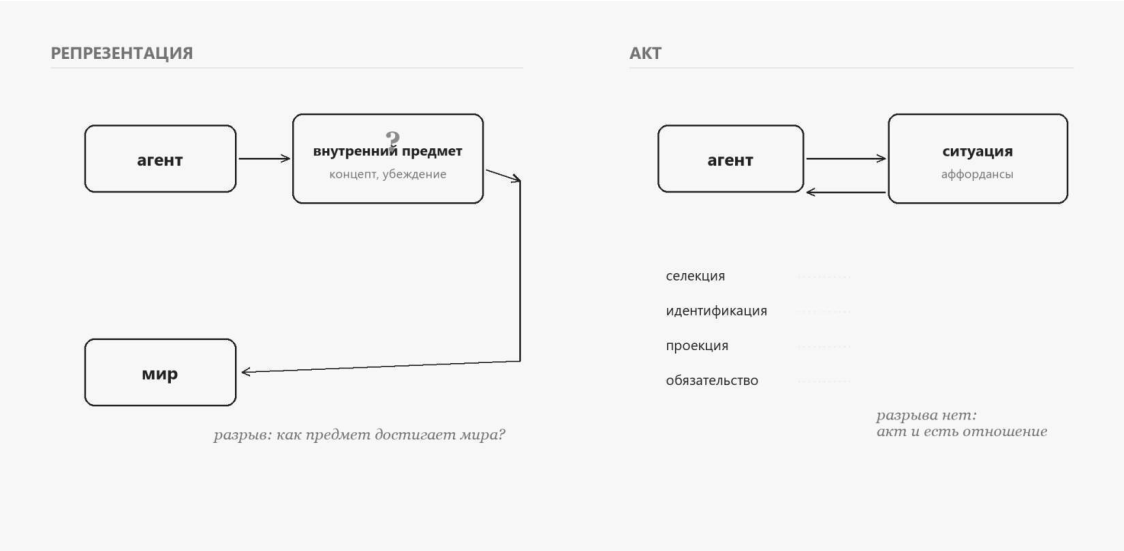


Рис. 1. Две картины познания: репрезентационистский разрыв и акт как само отношение.

Прежде чем двигаться дальше, нужны три уточнения, потому что эту позицию легко спутать с соседними, которых она не принимает.

Первое: первичность акта — это не бихевиоризм. Бихевиоризм полагал, что ментальный словарь должен быть заменён описаниями стимула и реакции. Утверждение здесь иное: акт не есть реакция на стимул и не наблюдаем в бихевиористском смысле, потому что релевантная единица включает селекцию того, что считается ситуацией, идентификацию того, что она предоставляет, и проекцию последствий, которые так и не наступают. Наблюдаема структура успехов и неудач в разных ситуациях; акт есть организующий принцип, который делает эту структуру понятной. Сказать, что акт первичен, значит сказать, что структуру нельзя разложить на пары «стимул — реакция», не утратив ровно того, что её объясняет.

Второе: первичность акта — это не инструментализм. Прагматистская традиция иногда говорит, что понятия суть инструменты, а инструменты оцениваются по тому, работают ли они. Это близко к моему взгляду, но не тождественно ему. Инструмент всё ещё остаётся предметом, внешним по отношению к деятельности, которая им пользуется; понятие же, с моей точки зрения, не используется актом, а есть модус акта. Разница обнаруживается в трактовке истины. Для инструменталиста понятие истинно, если оно полезно; для меня релевантное понятие — не полезность, а инвариантность: сохранение одной и той же компетенции при расширении практики. Понятие может быть полезным и не пройти испытания на инвариантность, а может быть бесполезным в данном контексте и всё же его пройти.

Третье: первичность акта не есть форма элиминативизма. Я не утверждаю, что понятий не существует или что разговоры о понятиях следует заменить разговорами о нейронных разрядах. Я утверждаю, что категория, к которой принадлежат понятия, есть категория актов и диспозиций к действию, а не категория предметов. Это утверждение об онтологии, и оно оставляет достаточно места для обычного словаря понятий при условии, что этот словарь понимается как выделяющий стабилизировавшиеся компетенции, а не ментальную мебель.

С этими уточнениями аргумент книги можно изложить одним абзацем. Познание есть совершение триадического акта: селекция области ситуации, идентификация её аффордансов, проекция траектории последствий. Акт завершается обязательством, которое связывает агента с определённым миром, и именно связывание делает акт суждением, а не всего лишь внутренним событием. Поскольку акт временен и укоренён в ситуации, его успех измеряется не соответствием репрезентации, а инвариантностью компетенции при расширении практики. Поскольку акт есть связывание, он с самого начала имеет нормативную структуру, и этика есть изучение того, как акты предвосхищают целостность других агентов. Поскольку акт может совершаться небиологическими системами, ответственность должна определяться архитектурно, а вопрос о субъектности машины становится эмпирическим вопросом о наличии триады и связывания, а не метафизическим вопросом о сознании.

Остальная часть части I расчищает почву для этого аргумента. Глава 2 реконструирует традицию и показывает, что её провалы имеют один источник. Глава 3 обращается к двум областям, в которых репрезентационистская картина укоренена глубже всего, — к научному моделированию и к словесному отчёту — и показывает, что обе лучше понимать как практики ориентации, а не как передачи содержания. Глава 4 рассматривает самый совершенный репрезентационистский артефакт, когда-либо созданный, — большую языковую модель, — и доказывает, что её достижения и её пределы в точности таковы, как предсказывает диагноз.

Акт, действие, событие

Через эту книгу проходят три слова, и их нужно держать врозь: событие, действие, акт.

Событие — происшествие, которое можно описать, не ссылаясь ни на чей замысел. Мост рухнул в 4:12; стекло разбилось; цена упала. События суть единичности в смысле Дэвидсона: их можно считать, датировать, связывать причинно и описывать бесконечно многими способами. В понятии события нет ничего, что требовало бы агента.

Действие — фрагмент поведения, выделяемый намерением или описанием. Вопрос Энском «что вы делаете?» допускает несколько верных ответов сразу: двигаю рукой, подписываю договор, обязываю компанию. Действия суть события под описанием, и описание поставляет замысел агента. Действие может быть намеренным, не будучи рефлексивным, и может быть описано верно, не будучи понято тем, кто его совершает.

Акт в смысле этой книги — не третий пункт того же списка. Это процесс, которым ситуация становится определённой для агента, а агент — связанным с ней. Акт не есть происшествие, которому случилось нести значение; это условие, при котором вообще что-либо может нести определённое значение для кого-либо. Поэтому акты нельзя считать так, как считают действия, и вопрос «сколько актов вы совершили?» поставлен дурно.

У различения три следствия, и каждое важно для дальнейшего.

Во-первых, акт индивидуируется своим пределом, а не своими причинами. Два физически тождественных поведения могут принадлежать разным актам, если завершаются разными обязательствами. Разрез хирурга и удар нападающего могут быть одним и тем же движением и не быть одним и тем же актом, потому что миры, которые они конституируют, различны. Тождество акта есть тождество мира, который он делает определённым, и это тождество нормативно прежде, чем физично.

Во-вторых, акт как таковой не дан интроспекции. Интроспекция поставляет действия и их сопровождение — образы, слова, чувство усилия, — а акт есть структура, которая делает действие понятным. Вот почему словесные отчёты, о которых идёт речь в главе 3, суть реконструкции, а не наблюдения: отчёт описывает действие, а акт — то, чего отчёт схватить не может.

В-третьих, акт допускает степени интегрированности, но не степени реальности. Слабый акт — всё равно акт; он слабо связывает, легко пересматривается и узок по охвату. Но полуакта не бывает так, как бывает полудействие. Либо агент связан, либо нет, и связывание есть порог, отделяющий простую ориентацию от суждения.

Утверждение, что акт есть условие возможности, нужно отличать и от утверждения, что он есть причина. Кантовские условия возможного опыта — не события внутри опыта; это то, что должно быть выполнено, чтобы у опыта вообще были предметы. Акт в этой книге играет похожую роль с одной существенной поправкой: условия суть не фиксированная таблица категорий, а исторические компетенции, и именно историчность делает их проверяемыми. Условие, меняющееся вместе с практикой, можно исследовать через практику; фиксированную таблицу — нельзя.

Поэтому теория не есть версия эмпирической психологии и не соперничает с её объяснениями. Психология объясняет, какие действия при каких условиях происходят. Теория спрашивает, каким должен быть строй агента, чтобы его действия были о чём-то вообще. Два исследования сходятся на одних и тех же явлениях и отвечают на разные вопросы — так грамматика и история языка сходятся на одних и тех же фразах.

Ещё одно уточнение, потому что слово «акт» несёт иную тяжесть в другой традиции. Книга не употребляет «акт» как синоним «деятельности». Деятельность есть система актов, организованная мотивом; акт есть единица, расчленение которой делает систему возможной.

Отношение здесь — часть и целое, и целому посвящена глава 14. Глава 2 берёт ту традицию, которая разработала это различение тщательнее всех, и это не та традиция, на которую обычно ссылается англоязычная философия.

Цена операционального вопроса

Действующий вопрос не бесплатен. Он покупает объяснительную силу ценой определённого рода общности, и эту цену следует назвать до того, как аргумент пойдёт дальше.

Первая цена в том, что операциональный вопрос ставит теорию понятий в зависимость от теории практики. Если понятие есть стабилизировавшаяся компетенция, то не существует ответа на вопрос «что такое понятие воды?», независимого от практик, в которых понятие используется. Понятие воды у химика, у сантехника и у ребёнка — не три приближения к одному предмету; это три компетенции с разными профилями, разными диапазонами и разными мерилами правильности. Это действительная потеря для того, кто надеялся, что понятие можно задать раз и навсегда, и действительное приобретение для того, кто заметил, что спецификация всегда протаскивала практику под видом определения.

Вторая цена в том, что операциональный вопрос делает теорию понятий эмпирической в том смысле, в каком понятийный анализ таковой не является. Если понятие есть акт, то его наличие есть вопрос о том, что агент умеет делать, а вопрос о том, что агент умеет делать, решается наблюдением, экспериментом и клинической оценкой. Философ, задающий операциональный вопрос, не может разрешить его из кресла, потому что кресло поставляет лишь один случай — компетенцию самого философа, — а компетенция как раз и есть то, что под вопросом.

Третья цена в том, что у операционального вопроса нет естественного завершения. На вопрос о предмете можно было ответить, указав предмет; на операциональный вопрос отвечают, описывая практику, а практики разветвляются. Всегда есть ещё один контекст, ещё одно расширение, ещё один критик. Отсутствие завершения — не дефект метода; это структура самого явления, которое исторично до самого основания. Но это означает, что теорию понятий невозможно завершить так, как можно завершить таксономию, а требование завершённости — один из мотивов, поддерживающих жизнь картины предмета.

Эти цены стоит платить, и причина в том, что картина предмета платит их тоже, только молча. Всякое определение понятия как предмета предполагает практику его применения, и эта практика либо игнорируется, либо трактуется как ещё один предмет. Действующий вопрос делает предположение явным, а явность есть начало честности.

Есть и приобретение, которое стоит назвать, потому что оно не очевидно из критики традиции. Действующий вопрос делает теорию понятий непрерывной с теорией навыка, а теория навыка — та область, в которой лучше всего понятны явления обучения, мастерства и срыва. Традиционной теории понятий почти нечего было сказать о мастерстве, потому что мастерство не есть обладание большим числом пропозиций; это обладание более богатой и более расчленённой компетенцией. Как только понятие становится актом, мастерство оказывается центральным случаем, а не запоздалой мыслью, и философия понятий становится непрерывной с философией практики.

Глава 2 начинается с Аристотеля, потому что вопрос о предмете был его изобретением, и изобретением гениальным: он спросил, что делает вещь вещью того рода, к которому она принадлежит, и ответил формой. Всё, что последовало, — средневековая теория интенций, кантовский синтез, фрегевская пропозиция, прототип — есть вариация на ответ, а не отход от вопроса. Чтобы увидеть, почему вариация не могла удаться, нужно увидеть, что вообще делал этот вопрос.

Глава 2. Зеркало и его критики

Аристотель изобрёл этот вопрос, и изобретение было неотделимо от его ответа. В «Категориях» он спрашивает, что значит, когда нечто сказывается о подлежащем, и отвечает структурой: сущность, качество, количество, отношение, место, время, положение, состояние, действие, претерпевание. Вещь есть вещь того рода, к которому принадлежит, в силу своей формы, и форма — это то, что схватывает определение. Обладать понятием, с этой точки зрения, значит постигать форму — обладать в уме той же структурой, что организует вещь. Ум становится формой без материи, как рука становится инструментом.

Изящество этого решения не должно затемнять его обязательство. Понятие есть структура, присутствующая в уме и соответствующая структуре, присутствующей в вещи. Знание есть отношение структурного тождества между двумя элементами, одним внутренним и одним внешним. Всё, что следует далее в традиции, есть попытка сделать это отношение понятным, и история её провалов есть история философии.

Авиценна ввёл различение, которое определило средневековую и ранненововременную дискуссию: различение сущности, как она существует в вещах, сущности, как она существует в уме, и сущности, рассматриваемой самой по себе. Различение было задумано для защиты объективности знания от изменчивости познающих, но породило новую проблему: каков статус сущности, рассматриваемой самой по себе? Если она не в вещах и не в умах, она, по-видимому, нигде, и всё же предполагается, что именно она делает и вещь, и мысль определёнными. Проблема не случайна. Она возникает потому, что теория нуждается в третьем царстве, чтобы объяснить, как две единицы могут быть структурно тождественны, притом что ни одна не причиняет другую как то, что она есть.

Кант увидел эту трудность и преобразил её. Понятие больше не структура, извлечённая из вещей; это правило синтеза многообразного. Категории — не описания мира, а функции суждения, и единство предмета есть единство синтеза, который совершает субъект. Это подлинный шаг вперёд, и это самое близкое, к чему традиция подходит к позиции, защищаемой в этой книге. Но Кант платит за этот шаг новым дуализмом. Синтез требует многообразного, данного независимо от правила, а правило требует субъекта, который его применяет. Результат — знаменитая пропасть между понятийным и интуитивным и вместе с ней постоянная угроза вещи самой по себе, которая никогда не бывает схвачена понятиями, организующими опыт.

Посткантианская традиция испробовала три выхода. Гегель растворил вещь саму по себе в саморазвитии понятия, так что структура мысли и структура действительности становятся тождественными в процессе самопознания духа. Фреге восстановил объективное третье царство смыслов, очищенных от психологии, чтобы пропозиции могли быть разделяемы, а истина — независима от любого субъекта. Феноменологическая традиция от Гуссерля до Мерло-Понти вернулась к переживаемому акту конституирования и попыталась описать его без аппарата трансцендентальной психологии.

Каждая из этих позиций серьёзна, и каждая защищалась с великой изобретательностью. Но обратите внимание на то, что у них общего. Во всех случаях понятие есть нечто, что стоит в отношении к миру: оно тождественно форме вещи (Аристотель), или есть синтез многообразного (Кант), или есть момент саморазвёртывания духа (Гегель), или есть смысл, определяющий референцию (Фреге), или есть интенциональная структура сознания (Гуссерль). Отношение — это то, что подлежит объяснению, а понятие — один из членов отношения. Вопрос всегда в том, как этот член прикреплён к миру, и ответ всегда требует дальнейшего отношения, которое требует дальнейшего члена.

Витгенштейн диагностировал эту схему с непревзойдённой экономией. В «Философских исследованиях» он показывает, что любое правило можно соблюдать бесконечно мно-

гими способами, а попытка задать правильный способ, добавив ещё одно правило, порождает регресс. Регресс не есть дефект какой-то одной теории; это судьба любой теории, которая трактует понимание как обладание единицей, применение которой есть дальнейший акт. Средство, предлагаемое Витгенштейном, — не ещё одна единица, а практика: следовать правилу значит участвовать в форме жизни, и правильность следования обнаруживается в том, что люди делают, а не в ментальном сопровождении.

Диагноз верен, а средство неполно. Практика есть структура актов, и Витгенштейн очень мало говорит о внутреннем строении акта, который делает структуру возможной. Он говорит нам, что обоснование приходит к концу, что согласие в суждениях предполагается, что след правила — в практике. Он не говорит нам, что делает агент, когда выделяет случай как подпадающий под практику, что он в этом случае идентифицирует и как он проецирует последствия того, что трактует случай таким образом. Результат — философия, сокрушительная как критика и тощая как конструкция. Она расчищает почву, но не строит.

Та же неполнота поражает и натуралистические альтернативы. Теория прототипов, начиная с Рош, заменила определения градуированными структурами и показала, что категоризация чувствительна к типичности, контексту и задаче. Это было эмпирическое достижение первостепенной важности, и оно разрушило классическую картину понятий как наборов необходимых и достаточных условий. Но теория прототипов остаётся теорией единиц: прототип есть репрезентация, абстрактная или конкретная, а категоризация есть процесс, сравнивающий стимул с ней. Сравнение требует метрики сходства, а метрика либо фиксирована заранее (и тогда она не может объяснить чувствительность к контексту, ради которой теория и создавалась), либо задаётся задачей (и тогда прототип не выполняет никакой объяснительной работы).

Энактивизм, от Варелы и Томпсона до Ноэ и Хатто, сделал самый многообещающий шаг: он поместил познание в сенсомоторную деятельность воплощённого агента, а не во внутренние репрезентации. Восприятие, с этой точки зрения, есть не приём картины, а осуществление практического навыка, а содержание восприятия задаётся возможностями действия, которые предоставляет ситуация. Это гораздо ближе к позиции этой книги, и я буду не раз на это опираться. Но у энактивизма есть две слабости, которые настоящий аргумент призван исправить. Во-первых, он склонен оставаться на уровне сенсомоторных сопряжённостей и мало говорит о проективной, антиципаторной структуре, которая отличает понятие от простого навыка. Во-вторых, он часто довольствуется утверждением, что познание энактивно, не специфицируя нормативную структуру энакции — что делает акт правильным или неправильным и что связывает агента с исходом.

Прагматистская традиция от Дьюи до Брэндома восполняет часть того, чего не хватает энактивизму. Для Дьюи исследование есть трансакция между организмом и средой, а понятия — инструменты разрешения неопределённых ситуаций. Для Брэндома понятийное содержание есть выводная роль, а нормативная структура утверждения есть структура обязательств и полномочий. Оба правы в том, что понятие нормативно и практично. Но инструментализм Дьюи оставляет понятие истины недоопределённым, а инференциализм Брэндома оставляет отношение между выводом и миром неописанным. Инференциалист может сказать, что из чего следует, но вопрос о том, что ограничивает всю игру — почему одни ходы правильны, а другие нет, — решается указанием на восприятие, а восприятие как раз и есть самое тонкое место теории.

Теперь схема видна. Каждая крупная традиция выявила нечто верное: понятие структурировано (Аристотель), оно есть синтез (Кант), оно нормативно (Витгенштейн, Брэндом), оно воплощено (Мерло-Понти, энактивизм), оно практично (Дьюи), оно градуировано (Рош). Чего ни одна традиция не сделала — не дала единого описания акта, в котором все эти черты суть моменты, а не отдельные обязательства. Понятие структурировано, потому что у акта есть фазы; оно есть синтез, потому что фазы интегрированы в одно временное единство; оно нор-

мативно, потому что акт завершается связыванием; оно воплощено, потому что акт укоренён в поле аффордансов; оно практично, потому что акт есть способ совладания с ситуацией; оно градуировано, потому что фазы допускают степени определённости и успеха.

Утверждать это не значит заявлять, что традиция просто заблуждалась. Это значит утверждать, что традиция задавала вопрос, грамматика которого вынуждала её распределять черты акта между отдельными сущностями: формами, правилами, смыслами, прототипами, практиками. Именно это распределение и производит знакомое противостояние реализма и идеализма, объективизма и релятивизма, данного и сконструированного. Если черты принадлежат одной вещи — акту, — противостояние теряет хватку, потому что альтернативы были порождены расщеплением этой вещи.

Деятельностная традиция

Есть традиция, которая ставит версию этого вопроса уже сто лет, и это не англоязычная традиция. В советской и постсоветской философии деятельности единицей анализа никогда не была репрезентация — ею было дело: предметное действие и система действий, организованная мотивом.

Традиция начинается с Выготского и его учения об опосредовании. Ребёнок не сталкивается с миром напрямую; он сталкивается с ним через орудия и знаки, и орудия суть не внешние подспорья, а условия возможности высших психических функций. Утверждение, что слово есть орудие, — не метафора об общении, а тезис о строении познания: знак перестраивает акт, и перестройка делает акт специфически человеческим.

Рубинштейн дал онтологическую формулу: сознание и деятельность не две субстанции, а две стороны одного процесса. Психика не есть вместилище, которое затем выражается в поведении; она формируется в деятельности и существует как регулирование деятельности. У формулы тот же противник, что у тезиса этой книги, — картина ума как вместилища, и она приходит к близкому выводу другим путём.

Леонтьев разработал строение. Деятельность организована мотивом, действия — целями, операции — условиями. Три уровня суть не иерархия причин, а иерархия описаний: одно и то же поведение может быть операцией, действием или деятельностью в зависимости от того, чему оно подчинено. Интериоризация превращает внешнее действие во внутреннюю операцию; экстериоризация совершает обратное превращение. Эта схема — самая разработанная в традиции картина фаз дела и ближайший предок триады, защищаемой здесь.

Ильенков добавил онтологическое измерение, которое критики традиции обычно не замечают. Идеальное, по Ильенкову, не есть мысленное содержание; это форма деятельности, опредмеченная в вещах и воспроизводимая общественным субъектом. Орудие идеально в этом смысле: его форма есть форма деятельности, которая его создала и которую он делает возможной. Утверждение предвосхищает понятие аффорданса и идёт дальше, потому что делает объективность идеального вопросом воспроизводства практики, а не соответствия.

Щедровицкий и методологическая школа обобщили схему до систем деятельности с прямой целью проектировать и критиковать институты. Категория акта деятельности служила для обозначения единицы, у которой есть начало, предел и носитель, а методология применялась к образованию, проектированию и управлению. Слабость школы была обратной стороной её силы: схемы стали настолько общими, что описывали всё, а описание, подходящее ко всему, перестаёт различать.

Настоящая теория ближе к этой традиции, чем к любой англоязычной, и тремя вещами ей обязана. Первое — первичность предметного дела: единица не состояние, а делание, и делание направлено на предмет, который конституируется в самом делании. Второе — конститутивная роль орудий: инструмент не подспорье для уже готовой компетенции, а условие её возможности. Третье — общественный характер идеального: объективность воспроизводится, а не даётся.

Есть и четыре расхождения, и именно они составляют вклад книги.

Первое: традиция описывает фазы дела, но не выделяет фазу, которая его завершает. Мотив, цель и операция суть уровни организации; ни один из них не есть связывание, делающее дело притязанием на мир. Схема Леонтьева способна описать дело, безупречно организованное и никогда не совершённое, и категории для этого различия у неё нет. Четвёртая фаза — добавление.

Второе: традиция генетична и описательна там, где настоящая теория нормативна и проверяема. Она говорит, как складывается компетенция; она не говорит, когда компетенция объ-

ективна. Критерий инвариантности при расширении — добавление, и именно он делает теорию отзывчивой к свидетельствам, а не к истории развития.

Третье: традиция трактует субъекта как сплошь общественного и почти не оставляет места субъекту, индивидуированному функционально. Для её задач это не дефект, для наших — ограничение, потому что вопрос о машине требует понятия субъекта, не предполагающего общественной биографии. Функциональное определение агента — добавление.

Четвёртое: категория идеального сильна и недоопределена. Идеальное у Ильенкова объективно потому, что воспроизводится, но само воспроизводство не получает проверки. Критерий инвариантности превращает догадку в критерий, и это превращение книга вносит в традицию, а не заимствует из неё.

Ничто из этого не есть опровержение. Это наследство с процентами: традиция построила онтологию дела, а книга пытается достроить конструкцию, добавив фазу, которая связывает, критерий, который проверяет, и субъекта, который может быть машиной.

Кантианское наследство и его долг

Кант заслуживает более длительного разбора, чем позволяет реконструкция, потому что настоящая теория ближе к нему, чем к любой другой фигуре традиции, и её несогласие с ним соответственно острее.

То, что Кант увидел, а картина предмета затемняет, состоит в том, что единство предмета не дано, а произведено. Многообразное созерцания ещё не есть вещь; оно становится вещью через синтез, который совершает рассудок. Это подлинное предвосхищение утверждения, что акт конституирует свой предмет, и это причина, по которой настоящую теорию можно назвать натурализованным трансцендентализмом: условия возможности опыта — не фиксированная таблица категорий, а структура практики.

Чего Кант не мог увидеть, потому что рамка этого не допускала, — что синтез не есть правило, применённое к многообразному. Если бы это было так, регресс правил вернулся бы на уровне применения: правило нужно применить, а применение должно направляться дальнейшим правилом. Собственное решение Канта — схематизм — есть попытка остановить регресс, опосредствуя понятие и созерцание временной структурой. Попытка поучительна, потому что временная структура — именно там, где настоящая теория помещает акт. Схема есть самое близкое, к чему критическая философия подходит к триаде, и не случайно три фазы у нас временные, а не логические.

Долг перед Кантом, таким образом, двойной. Теория наследует прозрение, что объективность конституируется, а не даётся, и наследует проблему, что конституирование требует материала, который сам не конституирован. Настоящая теория платит этот долг, отказываясь от второго наследства: нет неконституированного многообразного, потому что ситуация всегда уже структурирована историей агента, и это структурирование и есть то, что совершает акт. Отказ этот не есть опровержение Канта; это уточнение того, чего требует его прозрение, как только дуализм понятия и созерцания оставлен.

Гегель увидел дальше и заплатил за это тёмностью. Если конституирование не укоренено в материале, независимом от него, то конституирование нужно понимать как процесс, порождающий собственные предпосылки. Это логика диалектики, и её цена в том, что у процесса нет внешней проверки. Настоящая теория принимает гегелевский тезис о том, что условия опыта историчны, и отвергает утверждение, что история есть саморазвитие единого субъекта. Именно этот отказ позволяет теории быть плюралистичной, не становясь релятивистской: практик много, у каждой свои условия, и практики ограничены последствиями, которые не принадлежат ни одной из них.

Фреге — противовес, и отношение теории к нему самое деликатное. Достижение Фреге состояло в том, чтобы отделить содержание суждения от психологии судящего, и это отделение есть условие любого описания объективности. Настоящая теория сохраняет отделение, не сохраняя онтологию, которую Фреге на нём построил. Содержание — не третье царство смыслов; это структура компетенции, которая может быть разделяема, потому что может быть осуществлена. Разделяемость — не тождество внутренних единиц; это схождение практик, и это схождение и было тем, что третье царство пыталось гарантировать.

Деликатность в том, что фрегевский довод против психологизма сохраняет силу. Если содержание есть компетенция, то кажется, что содержание меняется вместе с агентом, а эта изменчивость угрожает объективности логики. Ответ настоящей теории в том, что изменчивость ограничена практикой, а практика нормативна: один и тот же вывод действителен для всякого агента, компетентного в практике, и эта действительность не есть факт о психологии агента. Ответ не фрегевский, но он делает ту работу, которую делал фрегевский, и делает её, не полагая царства, не имеющего причинного отношения ни к чему.

В этом обещание операционального вопроса, и в этом причина, по которой остальная часть книги конструктивна, а не только критична. Однако прежде чем начать строительство, нужно расчистить ещё один участок почвы. Репрезентационистская картина — не только философское учение; она встроена в две практики, центральные для наук о сознании: в практику научного моделирования и в практику словесного отчёта. Если эти практики предполагают картину, критика традиции остаётся чисто академической. Если нет, если их лучше понимать как практики ориентации, то путь открыт для эпистемологии, построенной на действии.

Глава 3. Модели, отчёты и иллюзия прозрачности

Две практики поддерживают репрезентационистскую картину действеннее любого философского аргумента. Первая — научное моделирование: считается, что модель есть репрезентация системы, а успех модели состоит в её соответствии системе. Вторая — словесный отчёт: субъекта спрашивают, что он думает или как он решил задачу, и ответ принимается за отчёт о внутреннем процессе, который субъект может наблюдать. Обе практики необходимы. Ни одна из них не требует картины.

Начнём с моделей. Стандартное описание гласит, что модель есть упрощённая репрезентация целевой системы, а её адекватность измеряется тем, насколько хорошо она представляет релевантные черты. Описание сразу же сталкивается с проблемой идеализации. Модель газа трактует молекулы как точечные массы; модель рынка трактует агентов как максимизаторов; модель нейрона трактует дендрит как линейный фильтр. Ни одна из этих моделей не представляет систему такой, какова она есть, а те, что представляют её лучше, часто менее полезны. Стандартный ответ состоит в том, что идеализация есть опускание того, что неважно для данной цели. Но этот ответ тихо меняет критерий: адекватность уже не соответствие системе, а адекватность для цели.

Как только входит цель, модель перестаёт быть картиной. Она становится инструментом ориентирования исследования. Карта — правильная аналогия, при условии, что мы не мыслим карту как картину местности. Топографическая карта не похожа на гору; она кодирует набор решений о том, что нужно знать идущему, и её адекватность обнаруживается, когда идущий приходит. Две карты одной и той же местности могут быть несовместимы в своих условностях и обе адекватны, потому что адекватность относительна практики, а не сходства.

Это не форма релятивизма, и различие важно для всего аргумента книги. Карта может быть неверной: она может поместить мост там, где нет моста, и идущий потерпит неудачу. Неудача объективна, и она объективна без всякого сравнения карты с местностью в целом. Терпит неудачу определённая антиципация — что, следуя этим маршрутом, идущий перейдёт реку, — и неудача регистрируется в последующих актах идущего. Объективность карты состоит в инвариантности её указаний при расширении практики: одна и та же карта должна продолжать вести правильно, когда идущий заходит дальше, когда меняется погода, когда ею пользуется второй идущий, когда местность обследуется другим методом.

Та же структура применима к научным теориям, и история науки даёт решающие примеры. Теория флогистона не была случайным набором ошибок; это была исследовательская программа с областью успешного применения, набором успешных предсказаний и связной онтологией. Она рухнула не потому, что была сравнена с реальностью и оказалась несовпадающей картиной, а потому, что её центральная компетенция — предсказание исходов горения и обжигания — не выдержала расширения практики, ставшего возможным благодаря количественным методам Лавуазье. Кислородная теория восторжествовала, потому что её компетенция выдержала: те же понятия продолжали вести правильно, когда газы стали взвешивать, когда был установлен состав воды, когда периодическая таблица переорганизовала область.

Мораль в том, что историю науки лучше описывать как историю компетенций, а не как историю картин. Теория есть стабилизировавшийся способ структурировать область так, чтобы антиципации становились надёжными. Её истинность — не отношение копирования; это инвариантность её указаний при расширении практики. Вот почему теории могут быть приблизительно истинными, не будучи картинами, и вот почему понятие приблизительной истины было так трудно анализировать на модели соответствия. Приблизительность — не вопрос степеней сходства; это вопрос диапазона, в котором компетенция продолжает работать.

Вторая практика, словесный отчёт, представляет ту же структуру в более интимном регистре. Стандартный метод когнитивной психологии, кодифицированный Эрикссоном и Саймоном, предполагает, что мышление есть последовательная переработка информации в рабочей памяти ограниченной ёмкости и что субъект, которого просят думать вслух, будет вербализовать содержимое этой памяти, не изменяя процесса. Это допущение не невинно. Оно требует, чтобы релевантное содержимое было доступно субъекту, чтобы оно было закодировано в вербальном формате и чтобы акт вербализации был прозрачен — то есть не менял того, о чём он отчитывается.

Ни одно из этих требований не выдерживает проверки. Требование доступности опровергается феноменологией умелого действия: компетентный носитель языка не имеет доступа к правилам грамматики, а компетентный клиницист не имеет доступа к выводной цепи, порождающей диагноз. Требование кодирования опровергается данными о том, что значительная часть информации, направляющей действие, не пропозициональна — она несётся в перцептивных различениях, моторных схемах и аффективных градиентах. Требование прозрачности опровергается реактивностью вербализации: описание лица ухудшает последующее узнавание, объяснение суждения меняет суждение, а требование назвать причину часто порождает конфабуляцию, а не отчёт.

Критика протокольного анализа у Эдера делает это с исторической точностью. Метод был изобретён в Вюрцбургской школе, усовершенствован Эрикссоном и Саймоном и подвергся критике с нескольких сторон: интроспекция не является надёжным источником информации о ментальных процессах; вербальное затенение ухудшает перцептивную память; а требование вербализации навязывает задаче структуру, которой у задачи не было. Вывод не в том, что отчёты бесполезны. Вывод в том, что отчёты суть акты реконструкции, совершаемые в ситуации, которая включает требование отчёта, и что их эпистемическая ценность зависит от того, для чего их используют.

У этого вывода есть следствие, выходящее за пределы методологии. Если отчёт есть реконструкция, то отчёт — не окно во внутренний процесс, а дальнейший акт того же рода, что и акт, о котором он отчитывается. Субъект, говорящий «я сравнил два варианта и выбрал более дешёвый», производит структурированное описание, которое делает его поведение понятным для аудитории, включая его самого. Описание может быть точным, но его точность — не точность фотографии; это адекватность ориентирования. Отчёт направляет антиципации аудитории — чего ожидать от субъекта дальше, как с ним взаимодействовать, о чём спрашивать, — и оценивается по тому, оправдываются ли эти антиципации.

Как только отчёты увидены так, эпистемология первого лица меняется. Авторитет отчёта субъекта — не авторитет наблюдателя с привилегированным доступом; это авторитет компетентного агента, описывающего практику, в которой он участвует. Это настоящий авторитет, и он подводит специфическим образом: когда практика описана неверно, когда описание служит цели, отличной от точности, или когда компетенция агента не сопровождается тем родом рефлексивного доступа, который описание предполагает. Эти провалы не умаляют авторитета; они ставят его на правильное место.

Обе практики сходятся в одном эпистемологическом принципе, который я назову принципом оперативной адекватности. Репрезентация — модель, отчёт, понятие, теория — адекватна в той мере, в какой направляемые ею акты остаются успешными при расширении практики. Принцип имеет три части, и каждая часть есть испытание.

Первое — испытание расширением. Продолжает ли компетенция работать, когда ситуация разрастается — новые случаи, новые масштабы, новые инструменты, новые агенты? Понятие, которое работает только для тех случаев, которые его сформировали, есть привычка, а не понятие.

Второе — испытание помехой. Продолжает ли компетенция работать, когда практика возмущается — когда задача погружена в другой контекст, когда агент отвлечён, когда меняются ставки? Компетенция, которая рушится под возмущением, никогда не была общей; это был маршрут, заученный для одного пути.

Третье — испытание артикуляцией. Можно ли передать компетенцию другим, подвергнуть её критике и улучшению? Понятие, которое нельзя артикулировать, приватно в том смысле, который делает его бесполезным для науки, потому что наука есть практика совместного исправления.

Эти три испытания заменяют отношение соответствия, и делают это без потерь. Всё, что соответствие должно было объяснить, — истину, объективность, прогресс, различие между знанием и мнением, — может быть объяснено инвариантностью при расширении, и объяснено лучше, потому что инвариантность проверяема, а соответствие нет. Цена в том, что знание становится существенно историчным: это всегда знание на данной ступени практики, и его притязание на объективность есть притязание на то, что выживет, а не на то, что совпадает.

Идеализация без модели

Философия науки создала богатую литературу об идеализации, и большая её часть предполагает, что проблема есть проблема репрезентации: как модель, искажающая свой объект, может тем не менее быть адекватной? Настоящее описание переформулирует проблему. Идеализация — не искажение, оправданное полезностью; это селекция области ситуации, достаточно богатой, чтобы поддержать антиципации, которых требует практика, и эта селекция есть тот же акт, который будет описан в главе 5.

Рассмотрим закон идеального газа. Закон не описывает газ; он описывает поведение класса ситуаций, в которых взаимодействия между молекулами пренебрежимы по сравнению с их кинетической энергией. Класс не вымышлен, и закон не приблизительно истинен в силу того, что ближе к истине, чем его соперники; он истинен для этого класса, и класс идентифицируется практикой использования закона. Идеализация есть граница класса, и граница обнаруживается нахождением случаев, в которых закон терпит неудачу.

Вот почему идеализация неотделима от измерения. Знать, где модель работает, значит знать, как измерить отклонение, а измерение есть акт, которого сама модель не поставляет. История газовых законов есть история инструментов, сделавших отклонения видимыми: термометр, манометр, калориметр. Без инструментов идеализацию было бы невозможно отличить от описания, а различие между моделью и её областью было бы невидимым.

Та же структура объясняет, почему одни идеализации плодотворны, а другие бесплодны. Плодотворная идеализация открывает диапазон случаев, в которых её можно проверить и расширить; бесплодная закрывает диапазон, делая свои провалы неизмеримыми. Различие не в внутреннем изяществе модели; оно в практике, которую модель делает возможной. В этом смысл того, что критерий главы 3 оперативен, а не только эпистемичен: он оценивает модели по практикам, которые они поддерживают, а не по картинам, которые они содержат.

Есть заманчивое возражение. Если адекватность относительна практики, то практика, отказывающаяся измерять свои провалы, будет производить модели, адекватные по умолчанию. Возражение верно и указывает на действительную патологию: практика, изолирующая себя от последствий, есть практика, которую нельзя исправить. Ответ не в том, чтобы отказаться от критерия, а в том, чтобы добавить условие: практика адекватна лишь в том случае, если она включает процедуры обнаружения собственного провала. Большинство практик этому условию не удовлетворяет, и то, что они ему не удовлетворяют, есть важнейший диагностический результат, который даёт теория.

Эту цену стоит платить, и остальная часть книги показывает почему. Триада главы 5 есть внутренняя структура акта, который предполагают три испытания. Обязательство в главе 8 — то, что делает испытания обязывающими, а не рекомендательными. А этика главы 11 — то, что происходит, когда расширение практики включает других агентов, чьи акты конституируют тот же мир.

Однако прежде чем обратиться к триаде, нужно преодолеть ещё одно препятствие. Репрезентационистская картина обрела новый дом, и он не в философии. Он в машине. Большая языковая модель — самый успешный артефакт, когда-либо построенный на допущении, что познание есть манипуляция репрезентациями, и её защитники и её критики почти ни в чём не согласны, кроме этого допущения. Глава 4 доказывает, что достижения модели подтверждают диагноз, а её провалы подтверждают альтернативу.

Глава 4. Идеальные репрезентационисты

Есть соблазн рассматривать большие языковые модели как опровержение репрезентационистской картины. Они производят беглый текст, отвечают на вопросы, решают задачи по математике и программированию, и при этом у них нет тела, нет сенсомоторной петли, нет мира. Если познание требует воплощённости, модели должны терпеть неудачу; а они не терпят неудачу — по крайней мере, не так, как предсказывали самые сильные версии энактивизма. Отсюда соблазн заключить, что модели показывают: репрезентации в конце концов достаточно.

Я хочу утверждать обратное. Модели — самые совершенные репрезентационисты, каких когда-либо строили, и их конкретные провалы — ровно те провалы, которые предсказывает репрезентационистская картина, хотя и не в том виде, как их обычно описывают критики. Дело не в том, что модели «не понимают» в каком-то расплывчатом смысле. Дело структурное: они оптимизируют отношение между токенами, а отношение между токенами — не отношение между актом и его последствиями. Каждый достигнутый ими успех есть успех внутри замкнутости текста; каждый их провал есть провал на границе этой замкнутости.

Начнём с того, что модели делают. Языковая модель обучается предсказывать следующий токен по контексту. Цель чисто статистическая: максимизировать правдоподобие наблюдаемого продолжения. Ничто в цели не отсылает к миру, к агенту или к последствию. Обучающий корпус есть запись того, что было сказано и написано, и модель усваивает регулярности этой записи — включая, что примечательно, регулярности, кодирующие структуру мира, поскольку запись была произведена агентами, совладавшими с миром. Модель есть, по сути, сжатое резюме следов, которые успешное совладание оставляет в тексте.

Это не тривиальное достижение, и оно объясняет, почему модели кажутся рассуждающими. Структура мира присутствует в записи, и достаточно богатая статистическая модель может эксплуатировать эту структуру, чтобы производить правильные продолжения. Когда модель решает текстовую задачу, она решает её не так, как человек; она производит продолжение, принадлежащее распределению правильных решений. Различие невидимо в выводе и решающе в механизме.

Решающее различие обнаруживается на границе. Проблема заземления символов была сформулирована Харнадом для формальных символьных систем, и её урок в том, что символьная система, символы которой определены лишь их отношениями к другим символам, не имеет внутренней связи ни с чем вне системы. Коннекционистские модели должны были решить проблему, обучаясь репрезентациям на сенсорном входе. Языковые модели решают другую задачу: они обучаются репрезентациям на тексте, который есть символьная среда, и потому наследуют проблему заземления в новой форме. Символы теперь — векторы, но векторы по-прежнему определены своим положением в пространстве совместной встречаемости, а совместная встречаемость есть отношение между символами.

Стандартный ответ таков: проблему решает масштаб — при достаточном объёме текста структура совместной встречаемости становится достаточно богатой, чтобы кодировать мир. Это отчасти верно и в основе своей недостаточно. Богатая структура совместной встречаемости может кодировать очень многое о мире, включая причинные регулярности, поскольку текст порождён агентами, действующими в мире. Но кодирование не есть способность действовать. Модель, усвоившая, что вода утоляет жажду, не может проверить это утверждение, не может быть застигнута неудачей и не может скорректировать своё поведение так, чтобы коррекция определялась исходом, а не текстом. Её «коррекция» — шаг градиента по функции потерь, вычисленной по токенам.

Вот почему провалы моделей так показательны. Они терпят неудачу в задачах, требующих удерживать согласованный мир на протяжении долгого взаимодействия; терпят неудачу в

арифметике, выходящей за пределы образцов обучающих данных; терпят неудачу в пространственных и причинных рассуждениях, которые нельзя восстановить из одного лишь текста; терпят неудачу, когда правильный ответ требует знания того, чего в корпусе нет. Это не случайные провалы. Это провалы замкнутости: модель компетентна внутри пространства сказанного и некомпетентна на границе, где ответ зависит от того, что не сказано, но тем не менее истинно.

Дополнение выборкой из базы и использование инструментов часто подают как решения. Это не решения проблемы заземления; это расширения замкнутости. Модель, умеющая вызвать калькулятор, — это модель, имеющая доступ к процедуре, выходы которой суть токены; процедура заземлена в мире, но отношение модели к ней остаётся отношением между символами. То же верно для модели, которая может запросить базу данных или управлять роботом: если цель модели не включает последствия её действий так, чтобы они могли пересматривать её компетенцию, заземление взято взаймы, а не принадлежит ей.

Обучение с подкреплением на основе обратной связи от людей (RLHF) делает этот пункт резче. Техника оптимизирует модель относительно модели вознаграждения, обученной на человеческих предпочтениях. Модель вознаграждения — прокси, а прокси сам есть репрезентация того, что люди одобряют. Результат — система, очень хорошо производящая выходы, высоко оцениваемые прокси, что не то же самое, что производство истинных или полезных выходов. Провалы выравнивания, документированные в литературе, — угодливость, игровое использование спецификации, взлом вознаграждения — не ошибки реализации; это предсказуемое последствие оптимизации репрезентации ценности вместо действия под ценностью.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.