Здравствуйте, Дмитрий Игоревич.

Начну с близких планов. Сейчас я прохожу специализацию по алгоритмам от Тима Рафгардена на курсере; правда, я немного застрял на третьем курсе из четырех, но к концу мая я планирую его пройти. Вот сегодня написал правильно алгоритм Прима на куче, который меня задерживал. После этого я бы хотел начать Advanced Machine Learning от ВШЭ и Data Science Specialization от John Hopkins University. По ним я бы хотел спросить, согласны ли Вы, что я получу, пройдя их, релевантный карьерный капитал для работы в сфере ML? Иными словами, согласны ли Вы, что правильным вложением времени было бы их пройти? Я посмотрел программу, много для меня интересного. Также я думаю, что прошел бы специализацию по биоинформатике от университета San Diego, потом что я нахожу механизм транскрипции-трансляции ДНК очень красивым и хочу узнать о нем и о других подобных процессах подробнее. Может, где-то пригодится.

На данный момент я довольно сильно уверен, что, чтобы принести как можно больше пользы человечеству при имеющихся у меня ресурсах, я хочу быть исследователем в области AI Safety. Карьерный гид 80000hours.org [советует](https://80000hours.org/career-reviews/) начать с получения PhD в области машинного обучения. В дальнейшем, по-видимому, я хочу работать где-то в MIRI, FHI, CHAI, DeepMind, OpenAI… Если у меня есть именно такая цель, то где бы вы посоветовали мне пройти магистратуру и куда пойти потом? Я знаю, что в Сколтехе есть магистратура Data Science, по описанию она мне нравится (особенно тем, что там фокус именно на технологии ИИ, а не на работу с большими данными). Другой вариант – МФТИ, там очень много магистерских программ по теме, и туда будет легко попасть, если стать победителем в олимпиаде «Я - Профессионал». Считаете ли Вы, что, быть может, было бы лучше искать что-то вне России? Где, по Вашему мнению, получать PhD? Делать ли перерыв между магистратурой и PhD и если да, то какой? Сегодня я нашел и прочитал интересный текст про PhD (прикрепил его к письму), Ваше мнение тоже хотелось бы услышать.

Миша Ягудин очень советовал начать погружение в AI Safety с [этого](https://intelligence.org/embedded-agency/) текста (Embedded Agency).

Он большой и очень обзорный, там говорится про проблемы, связанных с поведением рациональных агентов в среде. Я пока осилил только первую часть. У меня очень мало знаний в этих областях (как и у всех). Стал читать текст [An Introduction to Löb’s Theorem in MIRI Research](https://intelligence.org/files/lob-notes-IAFF.pdf). Этот текст мне не понравился, потому что он чересчур обзорный и нестрогий. Потом я прочитал One Decade of Universal Artificial Intelligence. Вот это мне очень понравилось, хотя строгих определений дано не было.

Кажется, что я бы хотел строить свой диплом вокруг текста Embedded Agency или какой-то другой статьи на похожую тематику, которая меня еще больше заинтересует. Как Вы думаете, писать ли мне диплом с Вами или с кем-то другим? Я уверен, за лето я больше буду разбираться в поле и лучше понимать свои интересы. Сейчас я читаю книгу Ника Бострома Superintelligence и «Гарри Поттер и Методы Рационального Мышления» (возможно, это не очень релевантно, однако ее написал основатель MIRI).

Помимо ссылок на источники с веб-страницы, где лежит Embedded Agency, у меня есть еще один [список](https://humancompatible.ai/bibliography) рекомендованных книг. Всего очень много. В ближайшие месяцы (годы?) я буду читать материалы из него и из Embedded Agency.

Во время магистратуры я хочу посетить MLSS, [AI Safety Camp](https://aisafetycamp.com/), школу [Deep Bayes](https://deepbayes.ru/) , [воркшоп](https://intelligence.org/ai-risk-for-computer-scientists/) MIRI. Также необходимо уже к концу магистратуры иметь какие-то публикации, чтобы взяли на PhD с большей вероятностью.

Про ИУП на четвертый курс я уже думал, там все более-менее понятно. Как Вы считаете, в каких местах мне стоит пройти стажировки? И если меня возьмут в ШАД, то стоит ли стажироваться? На мой взгляд, совмещение института, ШАДа и стажировки понизит качество усвоения материала во всех трех местах.

Рекомендательными системами в будущем я заниматься не хочу, потому что, на мой взгляд, они решают проблему, которая, не будучи решенной, не наносит никому вреда. А неаккуратное использование сильного ИИ – это экзистенциальный риск.