Понятно, что решение, которое может быть сделано с помощью слабого ИИ, не сильно отличается от случайного выбора. Такое решение может оказаться верным по случайности, но не будет осознанным. Чтобы принимаемое решение могло быть верным, оно должно быть этичным. Этичность - свойство чего-то мыслящего, то есть сильного ИИ. Следовательно, ИИ, которому мы можем доверить принятие решений, должен быть сильным.

Аргумент 1. Невозможность сильного ИИ.

Джон Сёрль рассуждает, что такой ИИ с сознанием сделать не представляется возможным. Задуманный нами ИИ - формальная система с инструкциями, а сами по себе формальные свойства не приносят понимания. Например, можно научиться сопоставлять куски текста (ответы) на незнакомом вам языке другим кускам текста (вопросам) того же языка по данным вам инструкциям, но вы так никогда не поймёте, что они означают.

Между такой манипуляцией текста и настоящим изучением языка есть очевидная для нас разница - мы получаем понимание во втором случае. Значит в нас заложены какие-то казуальные способности, продуцирующие интенциональность. Невозможно, что мозг с этими способностями инстанцирует какую-то программу, ведь всегда "...найдётся такой предмет, который инстанцирует эту программу, но всё же не имеет никаких ментальных состояний"

Т. е. сильный ИИ не может быть создан, а значит никакой ИИ принимать решения за нас не может.

Далее предположим, что сильный ИИ все же возможен и рассматривать будем только его.

Аргумент 2. Невозможность выбора системы ценностей для ИИ

Добавление этичности в сильный ИИ неразрывно связано с выбором для него системы ценностей. Проблема заключается в том, что систем ценностей существует огромное множество, и единственно верной нет (ведь тогда бы ее придерживались все люди). Если мы берем какую-то одну конкретную систему ценностей для ИИ, то все люди, которые этой системы не придерживаются, будут страдать от принимаемых на основе этой системы решений.
Если же мы берем две и более системы ценностей и пытаемся их как-то скомпилировать, то возникнет проблема внутренних противоречий (например, для кого-то высшей ценностью будет общее благо, а для кого-то - личное). Из-за этих противоречий ИИ в каком-то своем решении принесет

одним благо, другим - вред, то есть компиляция не решит проблему выбор одной единственной системы.

Аргумент 3. Неоднозначность интерпретации вводимых в ИИ данных

Предположим, что мы как-то справились с предыдущей проблемой (или тоталитарно выбрали одну систему, или придумали схему выбора, или закрепили за каждым человеком собственный ИИ с ценностями отдельного человека). Следующий этап - ввод этих данных в сильный ИИ.
Мы не можем заранее понять, каким образом ИИ интерпретирует эти данные. Как при наставлении маленького ребенка мы не можем понять, правильно ли он их понял (точнее, понял ли так, как мы задумывали), так и ИИ может понять очевидные для нас самих мысли совершенно по-другому. Например, мысль “необходимо заботиться об экологии”, которая в нашем понимании предполагает, среди прочего, переход к альтернативной энергии, сильным ИИ может быть воспринята как “необходимо уничтожить все, что экологии вредит”, т.е. человека. К тому же, мы не сможем понять, что именно привело ИИ к неправильной интерпретации.

Аргумент 4. Недетерминированность сильного ИИ

Продолжим аналогию из предыдущего аргумента. Так же, как мы не можем узнать помыслы ребенка, мы не можем узнать помыслы ИИ. Мы можем лишь попросить его объяснить причину какого-либо своего действия, но он может солгать. И мы никогда не узнаем, лжет он нам или нет.

Аргумент 5. Возможность глобального кризиса

Принятие решений ИИ может приводить к абсолютно разным последствиям, в том числе к последствиям глобального масштаба (гораздо более глобального, чем у любой существующей сейчас технологии). В том числе, и к наиболее катастрофичным.
Независимо от системы ценностей, ситуация, в которой можно как потерять все, так и получить сколь угодно многое, всегда будет неоправданной. Хоть ИИ и может крайне положительно сказаться на развитии цивилизации, тем не менее он может ее и уничтожить, то есть если ИИ будет принимать решения за нас, это может привести к утрате всего (хоть и с шансом вероятно крайне малым), и, соответственно, допущение этого неоправданно.

Вопрос другой команде: кто несет ответственность за принятое ИИ решение?