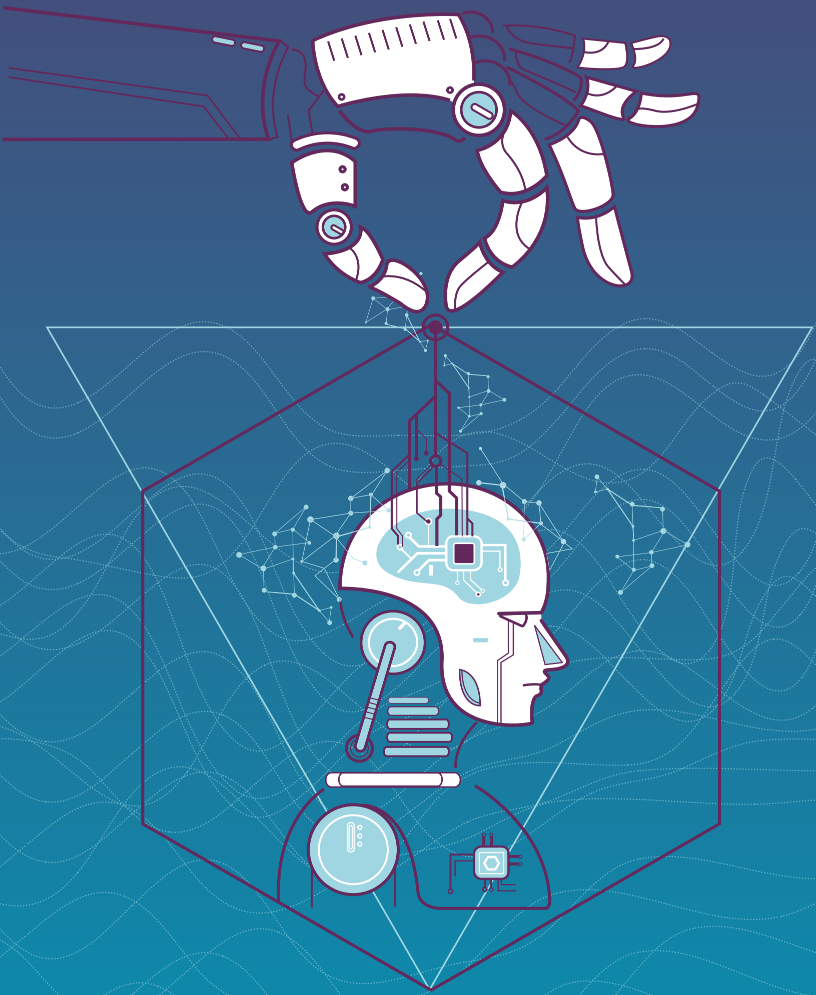
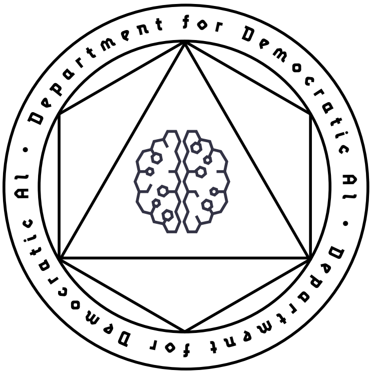
A DAY IN THE LIFE *of the* DEPARTMENT *for* DEMOCRATIC AI



HARRY FARMER



Department for Democratic AI • Department for Democratic AI •

A DAY IN THE LIFE *of the* DEPARTMENT *for* DEMOCRATIC AI

9am – Bernard Williams room

"So Jimmy, one final question." The young man opposite Julia shifted in his seat, preparing to pounce on this query as eagerly as the others she'd thrown at him that morning.

"What would you identify as the main challenges facing the Department?".

This one was always the clincher, Julia thought. As interview questions went, it was practically impossible to answer well: on the one hand, the Department didn't want to recruit anyone who saw its role as fundamentally flawed. But on the other, if you didn't have the wherewithal to see the difficulties – absurdities, even – of regulating the morals of AI systems, you really had no business as an adviser in one of the most scrutinised and challenging government departments.

"Well..." Jimmy began. "There are a few, actually. There are the problems faced by any body that exists to take morally charged questions out of the hands of politicians still in the firing line for them. Then there's enforcement. Because you can't realistically scrutinise every programme developed, you have to make the fines for non-compliance huge. It's impossible to make this sort of regulation light-touch."

"But most fundamentally, people's moral judgements aren't consistent. I don't mean they can't agree on answers to moral questions – though that also seems to be a bit of an open question." This aside drew a thin smile from Julia.

"I mean that it's seemingly impossible to come up with a set of general moral principles that, when applied to specific cases, consistently suggest courses of action we feel to be right."

For the first time in what had seemed like an eternity to Jimmy, Julia spoke.

"Some might suggest you've got the moral psychology backwards," she said. "They'd say that when people's moral intuitions clash with what follows from a moral principle they agree with, they're just as likely to dismiss their intuition as they are to abandon the principle."

Now it was Jimmy who smiled. He'd hoped this would come up today.

"So there's evidence this happens sometimes," he said. "But if you look at what happened a couple of years ago in the States – when people were asked to choose the moral principles of their cars – it's clear it doesn't always work that way."

Julia was all too familiar with what had happened when US regulators had decided it should be users of autonomous vehicles, not their manufacturers, who should decide how they should react to high stakes moral trade-offs. Every time you got in your car, it would ask you to reconfirm how you wanted it to behave: "before we set off, Mr Smith, I'll need you to confirm: would you like me to protect the car's occupants at all costs, or to minimise overall casualties and loss of life?". It hadn't gone well.

"Most people couldn't stand the idea of telling their car to respond to life or death situations in a particular way – not once they realised it made them responsible for decisions they could neither predict nor stomach. Those who got into crashes seemed pretty bad at rationalising their disgust with how the car had behaved – how they'd told it to behave. It didn't matter that the behaviour followed from principles they'd thought sounded about right."

"And why is this problem unique to the Department for Democratic AI?" Julia asked. "Why don't other institutions that regulate behaviour struggle with this? The legal system seems to more or less function, even though it also exists to codify and police norms."

Jimmy took longer before answering this one. He knew the answer – or thought he did – but he always struggled to express it clearly.

"Well, the difference is that the law only applies retrospectively. Of course there are legal principles that look forward – that allow people to understand what they are and aren't allowed to do – but sometimes we choose not to apply the law, to let people off the hook." Jimmy used a lengthy sip of water to gather his thoughts.

"It's different when the principles you decide upon are guaranteed to be followed to the letter, when there's no human to sense check or override them."

"The way we think about morals – I'd suggest – is based on specifics. But AI forces us to develop general principles, towards an uncomfortable level of absolution. That's a heavy burden for the Department to bear."

Julia smiled more fully this time. "Given these difficulties, do you think the department can do any good? And are there things we could be doing differently?".

The interview was more than twenty minutes over time by this point. Julia didn't really need to hear any more – she'd made up her mind to offer Jimmy the job a while ago. To her vague embarrassment, she was actually asking the question because she didn't know the answer herself. It wouldn't do for the Head of Policy to ask such a desperate sounding question of a recent graduate, but an interview provided the perfect cover for her to clutch at straws.

"I should say right away that I think what we have now is infinitely preferable to a world in which AI isn't regulated at all – or where regulators just look at outcomes rather than processes," Jimmy answered.

"If there's nobody keeping an eye on the decisions made by AI, then we're delegating, what? Eighty per cent of the choices made about us to something we can neither understand nor control. That's clearly not acceptable."

"For all the challenges, the Department at least provides a degree of certainty for those developing and subject to AI decisions. Even if the specific regulatory choices aren't always right, it's better to have imperfect ones than none at all."

Julia, who had been making notes on a tablet, looked up. "Thanks, Jimmy. We've overrun by quite a bit – but I think it was worth it."

2pm – Alan Turing suite

"Okay, let's try something different." Freya ran these workshops at least twice a week, but this one was proving tricky.

The thirty or so coders gathered around her were a rowdier group than usual. Most of those who attended her sessions were keen to get their machine ethics certification as quickly – and with as little effort – as possible, but this group had been disconcertingly engaged.

"I want you to break into groups again. I'm going to give you three different kinds of AI systems, and I want you to explore what moral issues you might encounter in developing them."

Slowly, the group shuffled itself into clusters.

"The first one should be nice and easy," she said. "A driverless car." A sigh rippled audibly across the room. "Yes, yes," Julia retorted, "but it's a cliché for a reason. We wouldn't be here if people had taken it seriously sooner".

"First, a driverless car system for use on mixed roads. Second, a system to optimise and monitor power usage in public buildings. Third, an assistive robot used by older people living alone."

"Before you start, I don't want you to try to solve any of the problems you identify. I just want you to come up with as many as you can, and tell me which you think are the most pressing. Any questions?".

Mercifully, and for the first time today, all Freya got was sullen stares. "Alright, I'm gonna give you half an hour. Then we'll go through them together."

With the task assigned, Freya grabbed her laptop and retreated to the corner of the room. While she enjoyed the subject matter – the idea that she could use her philosophy PhD outside of the academy still seemed too good to be true – she had recently been struggling to see the point in forcing coders to think about these things.

It wasn't as if she couldn't see the value of this part of her job. The government's requirement that anyone writing code in a professional capacity undergo ethics training was certainly an improvement on the past, when people would routinely develop software that turned out to have profoundly amoral dispositions – and then wring their hands when people noticed.

Nor was it fair to say that coders didn't care about the ethical implications of what they did. Ever since a landmark series of high court rulings established that it was developers of AI systems, not their users, who bore ultimate responsibility for their actions, most (though not all) programmers had been keen to cover their arses – though a lot had just found other, less fraught, careers.

The no-win no-fee AI compensation bonanza of the mid-twenties had taught everyone this the hard way. The party had been subdued when the company providing the algorithm used to identify potential claimants was itself sued into administration, and lawyers were forced to go back to drumming up business the old-fashioned way – with incredibly annoying adverts. It only truly ended when the Department had been formed and taken these cases out of the lawyers' hands for good.

The problem, Freya thought to herself, was that being aware of a problem wasn't the same as being able to fix it. As more than one of today's cohort had pointed out, it's all very well being aware of the moral dilemmas machines are likely to run into, and all very well – in principle at least – for a regulator to decide how they should respond.

The real problem was a technical one, bound up with the specific way most AI systems worked. As one particularly irritated coder had put it earlier this morning, "you're telling us to develop systems that follow particular moral principles. But you know that's not how AI works. Real AI is engaged with at the level of goals, not principles." Freya hadn't really had an answer to this, and struggled to get the group to see the point of their being there for the rest of the day.

Philosophers and cognitive scientists were working on solutions to this problem, the most promising of which was a project to develop 'explanatory systems' to run in parallel with existing AI – effectively allowing an AI to produce a rationalisation of its behaviour that coders could engage with in moral terms. But this was still a long way off. The difficulty was, more often than not, that the rationalisations produced by the AI were obviously obscuring the real reasons behind its behaviour. Freya, who had worked on this problem as a postdoc, suspected this was exactly the way the human brain worked, but until the machines got at least as good as people at pretending to think morally, it would be an issue.

Wary of the time, Freya forced herself from her reverie back to the task in hand. Glancing up periodically to make sure the coders hadn't mutinied in her absence, she began typing.

"Outcomes of notable recent decisions from DDAI's Citizen Councils – notes for Ministerial briefing…"

***

2:30 pm – Offices of the Secretary of State for Democratic Artificial Intelligence and Automated Systems

"Tariq, get in here now!"

"Yes, Minister?"

Tom had been Secretary of State for Democratic AI for almost six months, but still hadn't quite grown used to its permanent secretary's ability to appear silently within seconds of being called. Sometimes Tom entertained the suspicion (more seriously than was healthy) that Tariq was never really gone, just hiding somewhere, dormant, until needed.

"Have you seen the memo Freya just sent over – the one going over key changes to our regulatory principles since last year?".

"Of course, Minister."

"Well, how the hell am I supposed to justify some of these to the PM? I've got an interview with Theo Ashby from Youtube in four hours, what am I meant to say to him?".

"What specifically was bothering you, Minister?" Tariq asked. It was Tariq's job to understand the Minister's brief so he didn't have to, but half a year in, Tom's seemingly wanton ignorance of the nuances and paradoxes of the Department was becoming wearing.

The department was in an odd place. When it was founded, it was seen as somewhere a bright young politician could distinguish himself from his older peers, dealing with issues that were both unequivocally important and, so ran the prejudice, totally beyond the comprehension of anyone born before the nineties. The past two years had served as ample counterexample, but the Department was still new and shiny enough that a minister could pretend without too much determination that being sent there was an honour, not a punishment.

Tom was sharp enough to see the Department for what it was, but didn't quite have the self-awareness to accept he was unlikely to be leaving anytime soon – at least not for something better. The result was that he regarded learning the ins and outs of the place and its work as somewhat of a waste of time, knowledge he'd only have to discard once he'd been shuffled up to the Foreign Office or the Ministry of Resource Security.

"What's bothering me is that almost half of our new policies are totally inconsistent with our existing ones. One of the promises we made to industry when we set this place up – the main promise – was that even though the regulation we imposed would be onerous, we'd provide certainty. We said AI business would know where they stood."

"Until this morning, our position on AI paternalism – so carebots, personal avatar assistants, semi-autonomous exoskeletons, God knows how much else – was that a system can go against the stated wishes of its user if it's necessary to prevent clear and immediate physical harm to that person, or harm to others that would follow as a result of the AI's action – but not its inaction... I'm paraphrasing, obviously."

"Yes, Minister."

"So look at what she's just sent me." Tom gestured the text on his tablet up onto the wall and circled a paragraph. "She says this year's citizen councils have almost completely reversed this position. Assistive AIs basically can't intervene now – practically the only exception is that they can't help you to commit suicide."

"If I were the CEO of one of these companies, I wouldn't know where I stood. Hell, I'm the Minister of the department that makes the rules and I couldn't tell 'em where they stand. How do I justify this? We can't have our regulatory position change every bloody year."

"Well, Minister," Tariq began, carefully. "It's a different set of citizen councils to last year. They can't be expected to come to the same conclusions."

"I still don't get why the can't use the same bunch of people every year, or at least give them longer terms," Tom replied.

"The problem with that is that the councils are meant to be representative of the population as a whole. The mere act of serving on a council like this for any period of time is not a normal thing to do – it makes you less representative."

"But why is there such variation year on year? A bit of change I understand, but a one eighty pivot on such an important principle? The groups are meant to be pretty much the same, aren't they? If we choose them so carefully, how come there's this much variation?".

"We can't control for everything." Tariq said. "People's moral dispositions seem to vary in unpredictable ways."

"But surely we can. There must be correlations between the other data we collect and answers to the trolley problem? Why can't we recruit the councils like that?".

"Well technically, Minister we could –".

"– So?".

"But it would violate our own regulatory principles on data mining. And supposing we could understand people's fundamental moral dispositions in advance, then what? Would we pick council members on that basis? It would feel rather like loading the dice, don't you think?".

"Then there's the question of how you load the dice. Do you want equal representation between people with different moral psychologies, or do you want the fundamental moral psychology of the councils to reflect that of the country? And how do we even know what the country thinks?".

"This is all fascinating, Tariq, but how does this help me? We're still in a position where the rules are changing almost every year. It's just not acceptable."

"There's every chance things will settle down. If we know anything, it's that people aren't sure how they want AI to behave; these questions really are difficult. Right now, members have got very little to go on – AI morals have barely been regulated for four years now, and for the first two, nobody really knew about it."

"Future members will go in knowing full well what previous councils have decided. Given the huge levels of responsibility placed on them, by far the easiest thing for them to do will be to agree with what's come before. That way, if they get it wrong, they won't be the only ones. And the more this happens, the more likely it will be to happen. If five previous councils have decided on a set of principles, you've got to be damn sure of yourself to suggest something different. Give it a couple of years and you'll get your stable regulatory environment."

Tom pondered this for a moment. It would have been more comforting if he'd had any intention of being at DDAI for anything close to two years.

"That's all very well, but how is that meant to help me now? I can't say that tonight."

"I've prepared you some talking points that should buy you some time before this kicks in. They should be in your Red Box."

Tom glanced down at his tablet, opening up the Red Box folder. "Okay, I'll read this now. Thank you, Tariq."

"Of course, Minister."

Tom looked up to smile at his Permanent Secretary, but he was already gone, the door closed silently behind him.

***

nesta

This work was part of Nesta's Radical Visions
for Future Government collection, 2019

https://www.nesta.org.uk/report/radical-visions-future-government/