Society for Computers & Law
The leading educational charity
for the tech law community
www.scl.org

AI for Schools Programme

Notes for Presenters to accompany slide deck

Introduction

These notes are intended as an aid for presenters delivering seminars for the SCL AI for Schools programme. They indicate how the slides in the SCL AI for Schools slide pack are intended to be used and make suggestions as to how presenters may like to run their sessions in schools and what information should be covered. These notes are not intended as a script, but they do contain phrases, questions and approaches that you should feel free to use. References to Slide numbers are to the slides in the SCL AI for Schools pack. These notes are for the use of presenters only and should not be distributed to schools or students.

Note: some of the volunteers for the SCL AI for Schools programme are highly experienced presenters with existing in-depth knowledge of AI; others are not. Although we are aiming for broad consistency of delivery across the programme, more experienced presenters should feel free to disregard some of the more prescriptive suggestions in this note, which are intended as an aid rather than a restriction.

General preparation

- On the assumption that sessions will be delivered by presenters in pairs, split the slides between you and your partner.
- Read around the topics and gather some examples that are likely to be meaningful to your particular audience or that you will enjoy sharing with them.

Slide 1 – Title Slide

- [Nothing to say here! This is the holding page while your audience gathers.]

Slide 2 – Welcome and Introduction

- Introduce yourselves.
- Explain the structure of the session. For example:
    - *We're going to give you brief introduction to AI*
    - *We'll start by introducing AI and identifying some aspects of AI that make it legally difficult.*

- *Then we'll look at a few case studies which you may encounter if you become a tech lawyer to see what can go wrong.*
- *We want you to be thinking about whether the law works fine – or whether we need some new ones. What do you think the law should be?*

- Mention what isn't in scope. For example:
  - *There are a couple of really important AI legal issues we're not going to spend any time on – one is human rights; and the other is intellectual property rights.*
  - *Those subjects are really interesting and really important. But we can't cover everything – and today we're focusing on liability when things go wrong.*

Slide 3 – What is AI?

- ***Question:** Anyone want to have a go at defining it?*
- Don't be disheartened if none of the students want to put their hands up at this early stage. Just say something like '*Okay – let's have a look at what it is*' and move onto the next slide.
- If a student offers a definition, be encouraging – even if the definition is poor! Something like: '*I like that definition – really interesting. Let's see what the types of AI are and how they are typically understood*'.

Slide 4 – Strong AI

- Briefly introduce the concept of Strong AI. For example:
  - See e.g. https://www.ibm.com/cloud/learn/strong-ai
  - A machine showing human-level intelligence, able to learn and operate in different contexts. Currently, the stuff of science fiction – but how much longer?

Slide 5 – Weak AI

- Briefly introduce the concept of Weak AI. For example:
  - See e.g. https://www.ibm.com/cloud/learn/strong-ai
  - AI that focuses on a specific task.
  - This is what exists in the real world.
- ***Question:** Can anyone think of an example of 'Weak AI'?*

Slide 6 – Examples

- Pick up on any examples given in discussion.

- ***Question:*** *Would anyone like to explain one of the kinds of Weak AI shown on this slide?*
- Prepare to speak to a few of these examples. Medical imaging evaluation is compelling – see e.g. https://www.cnbc.com/2020/01/02/googles-deepmind-ai-beats-doctors-in-breast-cancer-screening-trial.html

Slide 7 – Machine Learning

- Explain that most AI available today involves machine learning.
- Describe what it is in simple terms. See e.g. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Slide 8 – Supervised Learning

- Briefly introduce supervised learning – see e.g. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
    - '***Supervised*** *machine learning models are trained with labeled data sets, which allow the models to learn and grow more accurate over time. For example, an algorithm would be trained with pictures of dogs and other things, all labeled by humans, and the machine would learn ways to identify pictures of dogs on its own. Supervised machine learning is the most common type used today*.'

Slide 9 – [Picture]

- This shows the challenge of very similar object recognition. Here: chihuahua vs muffin.
- Other illustrations include: labradoodle vs fried chicken; dog vs bagel; and sheepdog vs mop. See e.g. https://github.com/rcgc/chihuahua-muffin

Slide 10 – Unsupervised Learning

- Briefly introduce supervised learning – see e.g. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
    - '*In* ***unsupervised*** *machine learning, a program looks for patterns in unlabeled data. Unsupervised machine learning can find patterns or trends that people aren't explicitly looking for. For example, an unsupervised machine learning program could look through online sales data and identify different types of clients making purchases*.'

- Give an example, e.g. a recommender system for a streaming service like Netflix might use unsupervised machine learning to use viewing patterns to group users together and recommend other content.

Slide 11 – Reinforcement Learning

- Briefly introduce reinforcement learning - see e.g. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
    - '***Reinforcement*** *machine learning trains machines through trial and error to take the best action by establishing a reward system. Reinforcement learning can train models to play games or train autonomous vehicles to drive by telling the machine when it made the right decisions, which helps it learn over time what actions it should take*.'
- Expand on the walking example on the slide – programming an AI to walk would be an extremely complicated task. What if you put an AI agent into a simulated environment and created conditions so it learns to walk, then run, through trial and error?

Slide 12 – [Gifs]

- These are images of putting an AI agent into a simulated environment and getting it to learn to walk and run through trial and error. Taken from work by DeepMind.
- See https://www.deepmind.com/blog/producing-flexible-behaviours-in-simulated-environments
- Talk about how reinforcement learning can produce unexpected results when the AI 'games' the system.
    - See the discussion here: https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity
    - That DeepMind document links to the following list of specification gaming examples: https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml
    - Pick one or two – this robot hoover example is a good illustration: '*I hooked a neural network up to my Roomba. I wanted it to learn to navigate without bumping into things, so I set up a reward scheme to encourage speed and discourage hitting the bumper sensors. It learnt to drive backwards, because there are no bumpers on the back*.'

Slides 13-14 – Example: DeepMind

- Link back to the discussion about Strong AI by describing the progress made in developing general-purpose algorithms, i.e. DeepMind's MuZero masters Go, chess, shogi and Atari without needing to be told the rules: https://www.deepmind.com/blog/muzero-mastering-go-chess-shogi-and-atari-without-rules

Slide 15 – Some problems with AI systems

- Identify that there are some particular problems with AI systems. For example:
    - *They can be a 'black box' – with outcomes difficult to predict and explain.*
        - *AI systems can give nasty surprises!*
    - *They involve complicated technologies built from many parts by many people. If they go wrong, it can be hard to know what went wrong and why.*

Slide 16 – Case Study 1

- Explain that:
    - You will be discussing four different cases with the students. These are intended to give a flavour of the kinds of problems tech lawyers of the future may be advising on.
    - Each one starts with a short video setting up the scenario.
    - You will want to hear the students' views on some questions arising from them.
- Introduce the first scenario:
    - Tell the students to imagine they are designing a system for processing loan applications at a bank.

Slide 17 – [Embedded video]

- [This is an embedded video – make sure you know how to play it with sound!]

Slide 18 – Case Study 1: Predicting the outcome

- Introduce the first part of the Case Study:

  Start with the assumption that the system will not use AI – instead it will use a "deterministic algorithm" to decide whether to agree to the loan. Explain what that means (i.e. that the rules that will decide whether the loan is provided or not are set up front).
- ***Question:*** *What kind of rules would you build into that system?*
    - [Give them a couple of minutes to discuss amongst themselves.]

    - Explore with the students what kind of rules they would set in a deterministic system to assess applications.
    - Prompts if needed:
        - *Might check to see whether applicant has defaulted before*
        - *Might check that they have sufficient income to repay it*
        - *Might check they're over 18*
        - *Might check where they live.*
- ***Takeaway:***
    - *Whatever we choose to build into our algorithm, we always know how it's going to behave – for any application we give it, we can predict whether the loan will be accepted or not.*
    - *Easy to do – just run the application past all the checks that our algorithm carries out. If it fails any of the checks, we'll know why it did not get accepted.*

Slide 19 – Case Study 1: Predicting the outcome

- Introduce the second part of the Case Study:
    - Tell students to image they are designing another system for processing loan applications –this time using AI.
    - This time, there is no set of rules.
    - Instead, a **supervised machine learning system** will be trained for the task.
- Discuss how the system works. For example:
    - *We're going to feed it a whole bunch of old loan applications, and we're going to tell it in each case whether or not the borrower paid back to the loan on time.*
    - *The system then builds a model internally and, after a while, it can start to guess based on the information in the loan application whether the loan was paid back.*
    - *Once we're happy that it's getting it right enough of the time, we take the system live and use it for future applications.*
- Explain how the system is different from the earlier deterministic algorithm, focussing on the 'black box' problem. For example:
    - *The AI's behaviour in any particular case can't be predicted up front and can be hard to explain.*
    - *Whereas in the deterministic system, it is clear up front how any specific application would be treated, the only way to establish whether the AI system will agree to any particular loan is by putting the application through the system.*
    - *There are two reasons for that:*

- *[1] The lack of transparency – not possible to know <u>what it is</u> about the applications which makes the system either accept them or reject them.*
- *[2] It is a probabilistic system – there's no single thing or set of things that will make the algorithm decide one way or another; instead the decision emerges from the connections in the model. Essentially, the system is deciding whether the new application looks more like the set of applications from bad loans or the set of applications from good loans.*

- ***Question:** Why does it matter if I can't look at an application and work out reliably how it will be treated by the system?*
    - Prompts if needed:
        - *Might make it difficult to test: it's going to need to be validated with a lot of test cases before it can be let loose on a real system.*
        - *And even once you've tested extensively, you can't be sure that it's going to work well in the field.*
        - *You can't be sure that it won't occasionally give mad outcomes – come back to that.*
        - *You can't be sure it's going to be fair.*
        - *All of these things mean it can't just be set up and left: it's going to need constant checking and double-checking to make sure its decisions are ones that can be relied on.*

<u>Slide 20 – [Embedded video]</u>

- Show video to show how it pans out for Eliza.
- Comment on the outcome:
    - *Sounds ridiculous!*
    - *But explain that what happened here is a common situation with machine learning models: we can train them; we can then validate them. However, we can't see why it makes a decision in any particular case.*
    - *This means in this loan example we can't explain to the customer in any meaningful way why their loan has been rejected.*

<u>Slide 21 – Case Study 1: Predicting the outcome</u>

- ***Question:** Why does Eliza's problem matter?*
    - Prompts if needed:
        - *Feels very unfair*
        - *Can't help customer improve their chances next time*

        - *Lack of transparency means that there may be bias and discrimination in the system which hasn't been spotted.*
- Discuss how this problem could be addressed, e.g.:
    - *Using a simpler algorithm that <u>can</u> be explained.*
    - *Running variations of the application through the model to see what changes impact the decision.*
    - *Etc.*
    - [You may want to note that there is substantial ongoing research into transparent supervised learning approaches but that most current generation approaches suffer from this black box problem.]
- ***Takeaway:*** the mainstream machine learning systems by default are not transparent and decisions can't easily be explained.
- ***Question:*** *Show of hands – who thinks that transparency matters for loan applications?*
- Explain that transparency doesn't <u>always</u> matter:
    - Where AI is making important decisions affecting people's rights: like the bank example; say, a benefits decision; or a decision as to whether to hire staff – it matters a lot.
    - Where the AI's being used to suggest what music I might like to listen to next based upon what I listened to last, probably doesn't matter very much.
    - But it is always worth thinking about.

<u>Slide 22 – Case Study 2: The AI Lawyer</u>

<u>Slide 23 – [Embedded video]</u>

- [Play video.]

<u>Slide 24 – Case Study 2: The AI Lawyer</u>

- ***Question:*** *Should the operator of the AI be liable for the poor advice?*
    - [First discuss in groups. Give 1 or 2 minutes. Judge timing depending on extent to which the students seem to be engaging]
- ***Follow-up question 1:*** *Would the situation be different if the website <u>didn't</u> make clear that it was an AI tool and <u>didn't</u> make clear it was experimental?*
- ***Follow-up question 2:*** *Should the standard of the AI's advice be judged by reference to that of a competent lawyer, or a competent AI?*
- Discuss the current law building on the students' answers, e.g.: it would obviously depend on the details but:
    - *It's unlikely that the operator would be liable for the poor advice where it was clear that the tool was an experimental AI.*

    - *But if it wasn't obvious that the advice was being given by AI and a user would reasonably think that they were interacting with a human lawyer, then the operator would likely be liable.*
    - *As to the standard by which the operator would be judged: if the website made clear that it was an AI tool, then the question will be whether it was designed, trained and validated with reasonable care; if the website did not make clear it was an AI tool, the operator will likely be judged by reference to the output that a reasonably competent lawyer would have given.*
- ***Variation**:* Tell the students to imagine a change to the scenario:
    - *This time, the legal problem would be really difficult for a human lawyer to solve but the AI got the right answer easily.*
    - *The user goes to the website then thinks they had better get a real lawyer's advice.*
    - *They go to the lawyer, who gets the wrong answer (because it's just really, really hard).*
    - ***Question:** Should the lawyer by judged by the standard of the AI? Show of hands. Why?*
- Discuss the implications of this, e.g., with reference to doctors:
    - *Imagine a not too distance future where AI consistently gets much better results than very experienced doctors.*
    - ***Question:** Would you expect a doctor to rely on his own skill or should he use the AI tool to complement it? Show of hands. Why?*

Slide 25 - Case Study 3: The AI Journalist

Slide 26 – [Embedded video]

- [Play video.]

Slide 27 – Case Study 3: The AI Journalist

- ***Question:** Should the news site be liable for the defamatory statement of its AI bot?*
- Discuss the current law, i.e.:
    - General position – If you intentionally publish a defamatory statement you are liable even if you didn't realise it was defamatory.
    - There is a defence to defamation (Defamation Act 2013, s.3(4)) where the author expresses an opinion that an honest person could have held, but only the author actually held the opinion. An AI can't hold an opinion – it's not human. So the defence probably doesn't work.
    - Therefore the paper is probably liable. And would probably still be liable even if they had read it before publication.

- ***Follow-up question:*** *Would your answer be different if the new site operators had read the article before it was published and thought the opinion of the AI that the actor was an honest one?*
    - Then the honest opinion defence may be made out.

Slide 28 – Case Study 4: The AI Driver

Slide 29 – [Embedded video]

- [Play video.]

Slide 30 – Case Study 4: The AI Driver

- ***Question:*** *Should the pedestrian be able to sue the driver for not intervening?*
    - Discuss the practicalities, e.g.: Pedestrian can sue the driver. However, it may be difficult to succeed in practice as the driver will probably argue that they weren't negligent because nobody would have been able to avoid the collision given what had happened.
- ***Follow-up question 1:*** *Should the pedestrian be able to sue the car manufacturer for suddenly accelerating?*
    - *The pedestrian probably also has a claim against the manufacturer under the Consumer Protection Act 1987, essentially saying that the car is dangerous.*
    - *The car manufacturer will probably argue that as the true cause of the accident was a malicious act by a third party, it shouldn't be liable. But that probably won't succeed: the question will be whether the manufacturer should have been guarding against this sort of hacking attack, and suspect on our facts the Court would say it should.*
- ***Follow-up question 2:*** *Should the only party liable be the hacker who caused the accident?*
    - *There may well have a claim against the hacker.*
    - *But there may be lots of practical barriers to receiving compensation? Can you find them? Do they have any money? Etc.*

Slide 31 – What Next?

- ***Bonus scenario if there's time:***
    - Set out the trolley problem. Something like: *There's a runaway train careering down the tracks. If it carries on its current path it will kill five children who are playing on the track (which they thought was disused). You are standing by a set of points and you can see what's about to happen. You can also see that if you move the lever on the points you can divert the train*

down a different branch of the line, so it misses the children. But if you do that, it will kill an old lady who's picking flowers.

- ***Question:*** *What do you do?*
  - *You move the lever and you have deliberately killed the old lady.*
  - *If you don't move the lever you have stood by and watched five children die.*
- ***Follow-up question:*** *What has this got to do with AI?* For example:
  - *Potentially quite a lot.*
  - *There's an AI driverless car, driving in terrible weather conditions. Suddenly, another car, driven by a human, swerves out of a side road. Unless the AI takes avoiding action, it will crash, killing the occupants of the AI car and the other car; but the only avoiding action it can take is to avoid the pavement, thereby killing whoever is on it.*
  - *What should the car be programmed to do in those circumstances? Choose the least deaths? Choose the path that kills older people in preference to younger people on the grounds that fewer life years would be lost? Choose a path at random?*
  - *What about if the choice is between killing the AI car's occupants or killing a third party? Same problems arise.*

- ***Takeaway:*** Wrap up the session. For example:
  - *These are very hard questions.*
  - *Tell the students that after the seminar is over they should go away and think about who should be making these decisions: The designer of the AI car; the owner of the car; the user of the car; or should the approach be prescribed by law? And if there is going to be a law to decide, how do we get that law right?*

Slide 32 – Being a tech lawyer

- Please bring your own experience to this slide. Tell them about how you are connected to tech law.
- Techlaw.chat is a short podcast where the authors of these slides discuss scenarios like these.

Slide 33 – The Legal Profession wants Diversity

- This is a good opportunity to say everyone is welcome and the profession is working hard to improve diversity.

Slide 34 – Thank you