Ethics Scenarios Of Artificial Intelligence For Information And Knowledge Management And Library Professionals

Andrew Cox, University of Sheffield

Version 2 November 25th 2021

# Introduction

This document contains eight ethics scenarios about Artificial Intelligence (AI) relevant to Information and knowledge management and library professionals.

It builds on my recent report, Cox, AM (2021) The impact of AI, machine learning, automation and robotics on the information professions: A report for CILIP <https://www.cilip.org.uk/general/custom.asp?page=researchreport>

The document consists of ethics scenarios each followed by a set of notes which are prompts to discussion. The document ends with a set of summative questions, and a very select reading list.

By being made available in CC/BY/SA licence it is made possible for users to edit them to suit a particular sector or organisational context and to update them as new concerns emerge.

It is part of an on-going project to refine understanding of the ethical issues for information professionals, which will be published in the future.

There is already a vast literature on ethics of AI, but little (to our knowledge) specifically designed for information professionals as such.

In this context the purpose of the document is to enable information professionals to:

1) better understand AI and the roles they might play

2) think through some of the ethical issues

3) weigh up how existing codes of professional ethics apply, as well as their own personal stance.

The document follows my report in defining AI inclusively to encompass familiar and unfamiliar, already achieved and futuristic scenarios. This has the benefit of prompting users to recognise continuity with more familiar ethical dilemmas, as well as to reflect on what new sorts of challenge AI poses. It also seeks to represent the issues in the widest sense by:

* Reflecting the range of AI applications, including some relatively familiar ones
* Reflecting the range of the profession (e.g. those working across health, corporate and government information services, public and academic libraries)
* Considering responsible AI as not just as applying to the development process but to its ongoing use
* Considering the wider social impacts and dependencies of data and AI technologies
* Considering whether AI technologies merely reproduce existing social inequalities or actively promote social justice.

## Scenario 1: Supporting first responders

You are excited to be part of a team developing an AI driven tool to assist first responders in emergency health situations. The evidence from the pilot suggests that it will help making life-saving decisions leading to better outcomes in 90% of cases.

You are leading in the project on data collation and data quality. One of the requirements of the system will be to join up data about all citizens so that the AI knows enough about their medical history to enable the right response to be administered.

Excited you contact data providers in various health organisations to ask them for access to the data.

You are surprised when four of the data managers from local services demand a meeting. They say that they will refuse to supply the data, giving the following reasons:

1. Since no consent has been obtained for this data use it is unethical and illegal
2. It is an invasion of privacy to share the personal data
3. There is a security risk if the data is held centrally because it could be hacked.
4. It is unclear how the system works and they are worried that it will not benefit all victims equally, regardless of gender, social class or ethnicity. They are particularly worried that there is less benefit to those for whom there is limited health data, such as recent migrants.
5. There is no accountability of the system if it offers advice that leads to a mistake resulting in harm.
6. They are worried that the data might be shared or sold to a third party in the future.

You believe that the AI will save lives.

* What ethics and value issues do you identify?
* How would you do?

## Scenario 1: Notes

* AI systems need data. Therefore, information professional roles relating to AI are likely to be around data management or supplying data to a wider system (not just running AI systems for information services themselves).
* There is a dilemma in this scenario between a compelling benefit (saving human lives) and a number of ethical risks, which might or might not materialise.
* In addition, it is important to consider that in many domains some parts of this scenario might not be legal for data protection reasons.
* There is a need to differentiate legal, safety and ethical issues, though clearly they are inter-related.
* In what sense is it a violation of privacy simply to have data about oneself in another system? This might be dependent partly on exactly what the data is, who uses the system and how.
* Even if this is well accounted for, there is always a risk attached to joining up data if the system is compromised, hence the security concerns.
* A common problem with AI is a lack of transparency about how it works. This might be for a number of reasons. It could be because of the difficulties of explaining the technology, the complexity of systems involved or because machine learning was used to train the computer to make decisions rather than the criteria being written into the code by a human. Which forms of opacity can be resolved and how?
* Related to this is a risk of deresponsibilisation and lack of clear accountability when things go wrong.
* Recent news stories have raised doubts in the public’s mind about whether AI systems are biased.
* A final concern around health data could be commodification. Once data has been accumulated, how is it going to be used, shared or sold in the future?
* Health applications may have both compelling benefits and complex risks. They might be considered high risk areas of application. How do logics shift when AI is applied in less controversial areas?

## Scenario 2: Nudges

After the number of reported incidents of self-harm on campus has risen, your university has developed an app to help address issues around student mental health and well-being. The app offers a range of functions including customised mindfulness tips, alerts about well-being related events, and even links to healthy recipes. It can also generate incentives towards positive behaviours. For example, it will “nudge” students to move if they have been sitting for more than two hours in one place in the library by offering a discount on a drink in the cafe. The evidence suggests that this has a positive impact in encouraging students to move more, so improving their physical and mental well-being. Students effectively opt in to use of the app when they download it. The university says that by gathering data about student mental-health and well-being it is acquitting part of its duty of care towards students.

The library is being asked to share data about individual library usage with the app. In return it will have access to reports on user behaviour, which, it is suggested, will help improve customer service, as well as ensuring well-being. You are a bit concerned that there might be some ethical issues.

* What ethics and value issues do you identify?
* What would you do?

## Scenario 2: Notes

* The app is well meaning and seems to have many potential benefits (including access to information). Ultimately students choose to download it or not, so some level of consent is obtained.
* Information professionals have a longstanding commitment to privacy. Can this principle be balanced compromised to gain certain benefits and are the benefits promised here compelling? What specific data is being asked for?
* A related question might be about how informed this consent is. Is it an opt in or opt out system? Have clear rules been published about how the data is used and shared? Do students have access to the profiles that have been built about them?
* The nudging of behaviour seems quite problematic. Prompting people to move rather than stay seated for too long could well be beneficial, but is it ever right to try and change behaviour in this sort of covert way?
  + Does the wording of the message clearly indicate its nature?
  + Are there any commercial interests served by the café discounts?
* Further issues relate to what data about student well-being and mental health is being gathered, who will have access to it and how will it be used?
* How has well-being been defined? What training data has been used to teach the AI about “normal” behaviour?
* A fundamental question could be asked about how the app fits into a wider strategy around student mental health and well-being. Even if ethically designed, is technology the best way to approach this issue or does this need to be part of a wider strategy?

## Scenario 3: The voice assistant

A public facing information service has run a high-profile project to develop a voice assistant which can answer user queries. It is part of a campaign to help users understand AI better by promoting its use in a public setting. It is already proving popular and after a vote among users it has been given the name ANNA to make it as user friendly as possible.

The voice assistant handling user enquiries promises to release hard pressed staff to work on other duties. Furthermore, making users comfortable enough to ask sensitive questions has always been a problem at the open enquiry desk. It is thought that the public might be more willing to ask sensitive questions to an impersonal chatbot and it has been located in a sound-proof booth to ensure confidentiality.

It is not all plain sailing, however.

Some frontline staff are upset and say that the voice assistant is taking away one of the most satisfying parts of their job: interacting with the public to answer their questions. They say that this contact is very important to some users who have limited other human contact. They claim that the voice assistant simply cannot replace a human in the same role. They have refused to come to any training courses about the AI. There is also a rumour that the introduction of the voice assistant was linked to the decision not to renew some contracts for weekend staff.

There is another problem: the voice assistant struggles with many accents, especially those of anyone for whom English is not their first language. Users are very diverse and many were not born in the UK and are not native English speakers. The system supplier says that it is a long-term objective to widen the range of accents (and languages) that the assistant can understand, but there is not enough demand from their wider customer base for this to be an immediate development priority.

A third issue is that some have complained that calling the virtual assistant “ANNA” is sexist.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 3: Notes

* The voice assistant seems to offer many positive benefits in terms of access to information, promotion of interest in AI and AI literacy, as well as enhancing the image of the organisation (and the profession). One interpretation of staff objections is that they are simply resistant to change.
* This seems to be counter-balanced by some significant problems. An important one is the nature of the practice of the reference interview which emphasises that the presenting issue a user offers is often not the real issue that the user has. Reference interview practice helps elicit what the real information need is. It is hard to see how this might be reproduced by AI.
* AI in general is predicted to replace many jobs. This may not be as controversial if those jobs are very routine. But it seems that there could be an impact on human dignity if AI impoverishes work.
* There could also be seen to be ethical issues in how AI dehumanises roles both for the worker and the customer.
* There is also the problem that while the service benefits some people through the information it offers, including those with dyslexia, it is not usable by others, specifically those who might be thought most in need, such as people who are trying to operate in another language (such as refugees and recent migrants) or those whose speech is affected by a disability. In that sense its more useful to those who are already privileged. Would it be appropriate to offer some alternative service for marginalised groups, or is the starting point for the project wrong?
* It may influence your answer to reflect on the type of questions that the voice assistant might be answering. Is it answering service related questions such as opening hours or factual questions, or something more complex?
* Something else to consider is whether the voice assistant “remembers” conversations and hold extended interactions.
* Who has access to the data from interactions, just the library or the supplier of the technology too? Does it matter?
* Finally, the naming of AI like chatbots and voice assistants, which are in a service role, has been a point of controversy. If they are given female names this could be seen as promoting gender stereotypes.
* How far do you think the issues would be different if it was simply an online chatbot rather than a voice assistant?

## Scenario 4: A special collection

A private donor is offering a large sum of money for refurbishment of the archive centre. The offer is conditional, however, on prioritising digitisation and procurement of an AI based search tool to promote access to an archival collection of the personal records deposited by their historically influential local family. The collection contains a vast amount of unique material on many different topics of local interest. But it does include the letters of one member of the family who was a slave owner and others who held posts in the British Empire. Naturally, much of the language used in the texts reflects dated ways of thinking about gender, sexuality, ethnicity and disability. The donor insists on all the material being accessible as a condition of the bigger donation.

Some staff object to the whole project because they say it is telling a story about local history which is about privilege. Others say it should go ahead because it makes information easier to access and the prime duty of the service is access and preservation of knowledge.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 4: Notes

* The donation offers greater access to information in general and promotes the preservation of historically interesting documents within a specific collection. However, such special collections almost by definition reproduce past ways of thinking that we now may find deeply problematic. They reflect past privilege. Even priorities in digitisation, in retrospect, may reflect biases.
* What consultation processes might support the evaluation of the donation?
* What practices might be applied to describe the content in appropriate ways? Are these anything beyond what is needed for any historical collection?
* Does the library have an obligation to promote a fairer society, going beyond merely managing bias?

## Scenario 5: Forum moderation

A public facing information service has invested heavily in creating online communities for its users. These forums allow users to discuss anything related to their interests, from hobbies to politics. Surveys have consistently found that users consider the forums an invaluable source of information. They are seen as immensely beneficial in increasing engagement with the service, with an impact on levels of use of resources and customer interaction.

The problem is that sometimes discussions get heated, trolls create conflict and some users simply post unacceptable material. The forums have become so popular that hundreds of messages are being posted each day and it is now impossible to moderate comments manually.

The platform hosting the forums are offering an automated tool to undertake moderation. They say that it automatically blocks all unsuitable content. But they tell you it will block some harmless content too. Inquiring further it seems that it is particularly likely to block messages from activist groups. However, the platform say that the algorithm is based on unsupervised machine learning so it is not possible to explain precisely how it works.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 5: Notes

* The forums are offering not only access to information but a form of freedom of expression for users.
* There must be some duty of care to ensure that inappropriate behaviour and statements are dealt with. Is AI an appropriate way to do this? It could be quite likely that the algorithms used were developed in another context so do not see what is an inappropriate topic in the same way as an information professional might.
* How far do the following offer solutions?
  + Seeking to present the clearest possible explanation of how the moderation algorithm works, albeit that this is inevitably partial
  + Tweaking the algorithm to allow certain types of content with a risk of allowing some inappropriate content to be published
  + Combining the AI moderation with some level of human oversight or input (“a human in the loop”)
  + Creating a mechanism by which users can appeal against blocked messages.

## Scenario 6: The recommender system

Your new system uses AI to automatically recommend resources to users based on their past reading, searches and browsing history. As well as many complimentary emails, a number of messages have been received complaining about the system. Here are some examples:

*I know your surveillance system is watching my every move. But it’s my human right to search for whatever I want and read whatever I want. Who are my searches being reported to? Don’t tell me it is not monitored for “suspicious activity” because I don’t believe you.*

*I want to notify you that my search for “bombing” related to my history course where we are currently reading about world war two. I was quite shocked when I saw some of the resources recommended. Will this search go on my record? Please reply asap. I am quite worried.*

*Some of the library’s recommendations make no sense. Whoever is deciding on them clearly does not know anything about my subject.*

*I feel increasingly frustrated by your recommendation system. I think its misleading students about the most appropriate reading related to the carefully selected texts I set. Is it just the publisher just pushing their stuff? I seem to have noticed a bias towards X (publisher name). How does this recommendation work?*

*Your recommendation system is faulty. I have done a number of searches that should have shown up my book, but it has not come up once! Please rectify this error!*

*I keep getting wrong recommendations. Can I update my profile?*

*Yet another dumbing down of our education system. What about students having to think and then search for the book they need: rather than you handing them out the answer on a plate?*

*Your recommendations are systematically biased towards white male authors from the global north, writing in English. I am sick of your “recommendations”: they are just propaganda. A university library should represent the plurality of knowledge.*

*I do not remember consenting to being stalked by a librarian. It’s my business what I search for.*

*Am I able to opt out of the recommendation system? I do not want my data sold to a publisher.*

You believe that many of these fears are unfounded, but you feel troubled.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 6: Notes

* Evidently many users gain benefits in terms of access to knowledge from recommendation. Perhaps they outweigh the costs.
* Campaigners for freedom of thought have been troubled for some time about the chilling effects of systems that might be intended to be helpful, such as through recommendations, but create a sense of surveillance. This might have a differential effect on certain communities, e.g. those more marginalised in society, so exacerbating other forms of exclusion. The service might respond that it does not examine individual search profiles. But the existence of the data raises the possibility of situations arising where the state or another organisation asks for it.
* Transparency is an acute issue with AI, whose working is often rather opaque. Several of the emails reflect confusion about how the system works, including believing that the recommendations are made by a human. It is already hard for users to understand how some library search systems work and so understand the trustworthiness of results. AI means that it is increasingly hard to explain things like recommendations. What can be done to make the system more transparent?
* AI could reduce human autonomy and specifically in the information field reduce information literacy by making it “too easy” to find loosely relevant information and dampening people’s incentive to think more critically about information.
* Some of the comments suggest that the data might not be accurate. What mechanisms might be needed to update information to ensure accuracy?
* Is there an issue of informed consent?
* There may also an issue of bias. Recommendations based on a collection or patterns of use of a collection will reproduce biases in these. What if the collection already contains a bias towards certain “pale, stale and male” authors? Or if past usage patterns reflect older ways of thinking about a topic? This might be seen as a particular example of the creation of a filter bubble.
* One of the emails queries the potential sale of user data. The chances are the platform would be run by a third part vendor. Is it known how the data is used?
* Some of the emails suggest an impact on how the organisation and profession are viewed. Could this be an ethical concern?
* How far do the following offer solutions?
  + Requiring an opt in to receive recommendations
  + Qualifying all recommendations, e.g. by advising students to check with their tutors or including prompts to help users develop AI literacy.

## Scenario 7: Stakeholders

The chair of a group representing an important local community group phones you about the AI project you are working on. She sounds a bit upset. She asks why you are excluding her community from the project. Genuinely puzzled you ask her to explain.

She replies that she wants to know why her group does not have a representative on the steering committee for the project.

You assure her that as an information professional you can be trusted to consider all user needs, including those of the diverse communities served by the service. It simply is not possible to have every group represented directly, you suggest. Her community will be engaged via a planned survey.

She starts talking about how “white” staff in your organisation are. She says that “none of you” really understand her community.

You ask if she has specific concerns. She says they are particularly worried about the data used in the system to describe her community. She says that her community is not represented as they wish to be in the data.

You assure her that the data sources drawn on by the AI are produced by authoritative public bodies.

She says the community does not trust some of this data. Often the data is simply wrong, she says. Moreover, the assumptions built into the data are discriminatory. Past data represents past mistakes.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 7: Notes

* The main presenting issue here is around representation in the process of developing AI. We do not know the specifics of the AI in question, but many public facing systems have multiple stakeholders in society. How far should or can they be included in the process of developing and then governing the use of the AI?
* Professional expertise and neutrality is also being challenged here. It is at issue how far the profession can fully understand the needs of different user groups or rely on the trust of social actors that is able to do so. Given the lack of ethnic diversity in the make-up of the UK information profession this could be a concern.
* The specific concerns about data reflect a sociological understanding of the way that social inequality is reproduced. Many forms of social counting assume a norm. Groups who do not fit this norm are discriminated against or effectively rendered invisible. For example, the census may assume things about household structures that assumes a nuclear family “normal”, rendering others invisible or deprecated. By inscribing such norms into data forms of inequality are maintained. The management of data is not merely a technical issue, it requires thought about the fairness of how different groups are represented by data.

## Scenario 8: Project partners

Your manager asks you to represent the information service on a prestigious AI project involving a large number of external partners. Feeling rather proud to be chosen you go to the project kick off meeting.

The first speaker is the project lead, John, representing a well-known big tech company. You have read some rather damning things about his company in the recent news, but actually he gives a rather inspiring speech about the potential of the AI technology to improve lives. Pleasingly he spends a lot of time talking about the protection of privacy, but this seems to be his main ethical focus. This strikes you as rather narrow. You ask about the environmental impact of the computing resource likely to be used by the technology. John seems surprised by the question and says he “hadn’t really thought about that” but it could be looked at.

From the next presentation it appears that one of the companies supplying data to the project is making heavy use of clickworkers. You ask about how much these workers are paid. Again, the speaker struggles to answer. He says his expertise is in data integration and he does not really deal with data sourcing. He promises to ask colleagues.

Thinking about who is at the meeting you are also concerned about the representation of user groups. When you raise this, you are told that this is covered in the project plan by focus groups of users of a prototype in month 11. A related question you ask is about provision for people with limited digital access or skills: are the benefits of the AI system only for those who are digitally connected? Again, this does not seem to have been thought about.

You are beginning to be quite troubled about a number issues, but you are also aware that you are the only one asking these sort of questions. It strikes you looking around the room nearly everyone is a middle-aged, white male.

In the lunch break you ask John about how responsible operations is being handled in the project. He says his company’s ethics team are not represented directly on the project but will carry out a review of the initial design.

He doesn’t seem very happy by the line of questioning and you are worried that you are beginning to sound like a technophobe and that your questions might not reflect well on your organisation or the profession.

• What ethics and value issues do you identify?

• What would you do?

## Scenario 8: Notes

* Information professionals are likely to be involved in multi-partner projects to develop AI.
* It has been argued that within the Tech industry notions of ethics tend to be rather narrowly defined around certain topics such as privacy. Here the project seems to recognise that issue but many wider ethical issues are far less visible.
* Wider societal and environmental impacts of the AI industry are often ignored in statements of responsible AI that focus purely on the AI development process.
* Within Tech firms it seems ethics centres have been created but often have a marginal status in the organisation.
* Ethics considerations apply to AI tools throughout design, but it is argued that they need to continue beyond this into on-going use.
* Again, part of the problem of bias in AI relates to the limited perspectives of those who develop the software. Representation and consultation of user groups becomes a vital issue.
* One criticism of the project might be that it makes the assumption of technological solutionism. This sort of thinking assumes that technologies can unproblematically address what are really complex social problems. It also marginalises human skills and caring work. Do you think this could be considered an aspect of the issue here?
* How the information profession presents themselves within such collaborations is important. How can appropriate ethical doubts be raised without appearing technophobic?
* Since the scenario imagines you representing a particular organisation, can organisational benefits outweigh personal doubts?
* How far do the following offer solutions?
  + Continue to ask questions hoping to change the project from the inside
  + Accept the limits of the current project and try and make a strong contribution
  + Ask your boss to be relieved of these duties.

## Summative questions

1. How well do current codes of professional ethics help guide one through these scenarios?
   1. CILIP Ethical framework, https://www.cilip.org.uk/page/ethics
   2. IFLA Professional Codes of Ethics for Librarians, https://www.ifla.org/faife/professional-codes-of-ethics-for-librarians
2. Do we need to use more specific ethics codes, such as The ethical framework for AI in education (Institute for Ethical AI in Education, 2020) or the ODI’s Data Ethics Canvas or the NHS A guide to good practice for digital and data-driven health technologies (2021) which address many of the ethics issues?
3. Do we need state regulation rather than relying on organisations and individuals to follow ethics guidelines?
4. Does AI create fundamentally new challenges to the ethics and values of the profession? If so in what ways?
5. Ethics statements generally focus on the responsibilities of the individual professional, but what are the ethical challenges for the organisation?
6. How far do you think the focus should be on simply ensuring fairness within the status quo or should information professionals and services be actively promoting social good/ social justice? Is this a matter for professional or personal/societal ethics?

## Very select reading list

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review*, *37*(2), 60–68. https://doi.org/10.1177/0266382120923962

Cordell, R. (2020). *Machine Learning + Libraries: A report on the state of the field*. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf>?

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

Jones, K., & Salo, D. (2018). Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads. *College & Research Libraries*, 79(3), 304. doi:https://doi.org/10.5860/crl.79.3.304

IFLA. (2020). *IFLA Statement on Libraries and Artificial Intelligence*. <https://www.ifla.org/files/assets/faife/ifla_statement_on_libraries_and_artificial_intelligence.pdf>

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, 35(4), 957-967.

Padilla, T. (2019). *Responsible operations: Data science, machine learning, and AI in libraries*. *OCLC*. https://doi.org/10.25333/xk7z-9g97