

The Jisc logo is an orange square with the word "Jisc" in white, sans-serif font.

Jisc

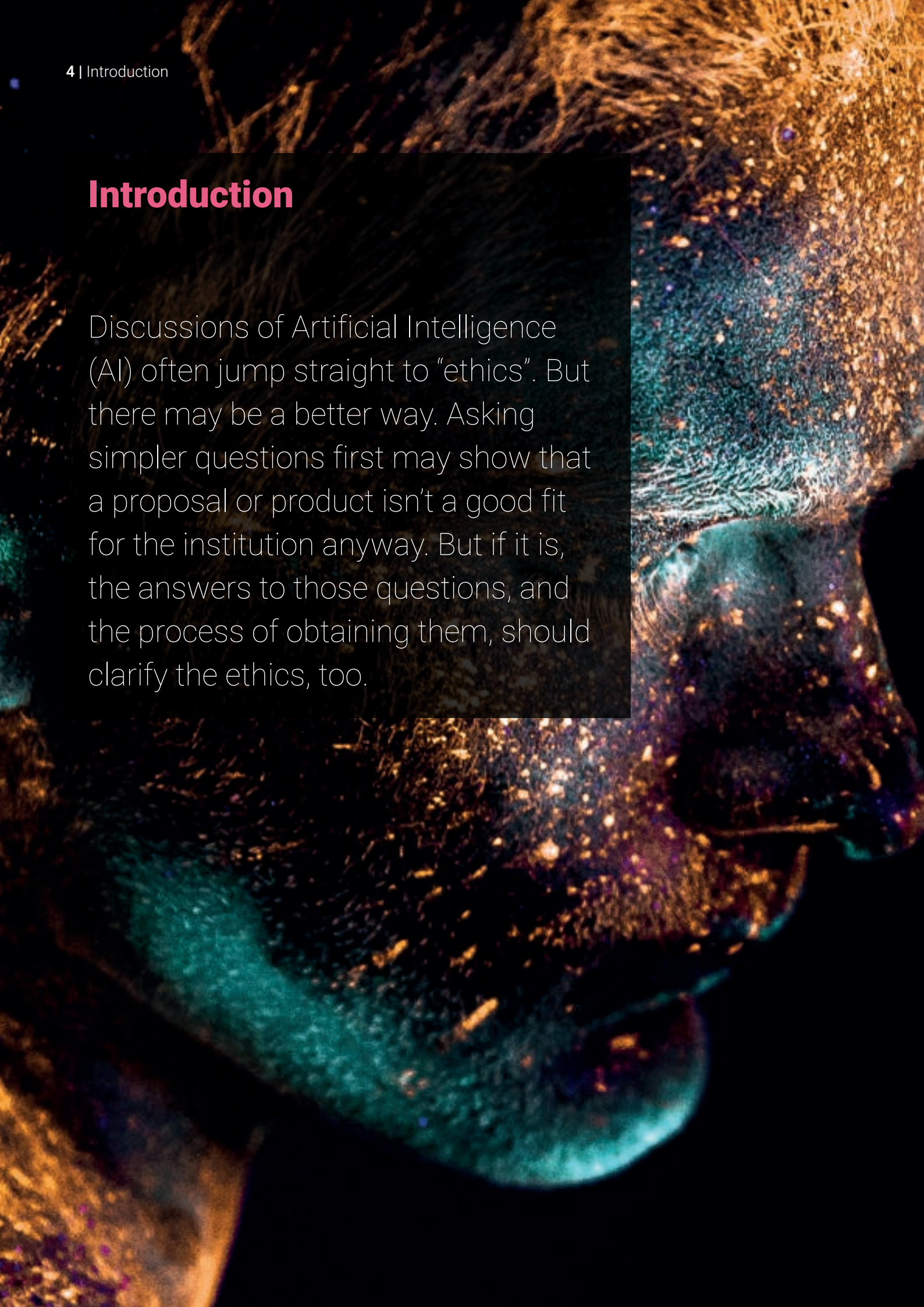
A pathway towards responsible, ethical AI

October 2021

-
- 4** **Introduction**
-
- 6** **Step 1: Does this proposal fit our institution's objectives?**
-
- 8** **Step 2: Does using AI fit our institution's purpose and culture?**
-
- 9** **Step 3: Are we ready to do this?**
-
- 10** **Step 4: Does using AI raise specific issues?**
-
- 13** **Finally: the details**
-
- 14** **Summary of pathway**
-
- 15** **Next steps**

Introduction

Discussions of Artificial Intelligence (AI) often jump straight to “ethics”. But there may be a better way. Asking simpler questions first may show that a proposal or product isn’t a good fit for the institution anyway. But if it is, the answers to those questions, and the process of obtaining them, should clarify the ethics, too.



This Guide suggests a pathway towards responsible, ethical AI with a series of discussions helping to quickly assess ideas and their fit for the institution. This won't cover the extensive literature on AI Ethics and doesn't focus on specific definitions of either "AI" or "ethics". Wherever an idea relies on data or algorithms it should help you think about "should we...?" as well as "can we...?".

Whether you are choosing, using, or benefiting from Artificial Intelligence, we hope this pathway will make you more confident in your relationship with it.

Using the Pathway

The pathway has four main questions:

1. **Does this proposal fit our institution's objectives?**
2. **Does using AI fit our institution's purpose and culture?**
3. **Are we ready for it?**
4. **Does using AI raise particular issues?**

Only the final stage considers possible technologies or implementations: here specialist guidance may be needed. Until then we focus on human and institutional acceptability.

The quickest way to assess whether an idea raises ethical concerns is to talk to the people it will affect: those whose data will be used, who will interact with it, and who – perhaps unknowingly – may be impacted by its decisions and recommendations. Discussions with minority groups are particularly important: AI may entrench existing unfairness or have unexpected effects. As a guide, we have suggested stakeholder groups that should be involved in each stage of discussions. Similarly, reference frameworks are ideas for how discussions might be structured. If you find other resources more helpful you are welcome to use them, and please let us know so we can update our references.

We hope you find this pathway useful.

Figure 1 Pathway Towards Responsible, Ethical AI



Step 1: Does this proposal fit our institution's objectives?

The first step is to discover whether an idea could deliver sufficient benefit to be worth taking forward. Break this down into three discussions, recording who was involved, what the outcome was, and any significant points of agreement or disagreement – this will be useful later on.



Why are we doing this?

A question for the idea's **proposers** and institutional **management**. Here we are seeking to confirm the rationale for the proposal: what is the anticipated outcome, how will that help the institution or its stakeholders, what benefits are expected? "Because technology/vendor lets us do it" is, at best, a weak rationale, and may be a sign that problems lurk further along the pathway. The Information Commissioner's **guidance on Accountability** (<https://ico.org.uk/for-organisations/accountability-framework>) is a useful framework, even for proposals that don't seem to involve personal data.

Will it work?

A broad question that needs discussing with **students, academics, tutors, professional services, implementers** and other **stakeholders** who may be directly or indirectly affected. It's surprising how many ideas fail if scrutinised from this perspective, so this is a good time to engage with critical friends and sceptics.

- Focus on dependencies: if the idea works, what then, do we have processes to use its outputs, for example?
- Will the idea generate the meaningful outputs those processes need?
- How will those affected feel about it, is it comforting, challenging or something they would seek to evade?
- What wider effects and assumptions are involved, what will happen if the environment changes, how might it affect different groups? As well as traditional equality, diversity and inclusion categories, discuss with those who learn from books rather than online, or have limited access to technology.

These four questions are discussed and illustrated in a paper "**Between the Devil and the Deep Blue Sea (of Data)**" (<https://doi.org/10.19164/jlitt.v1i1.1005>).

Does it advance our mission?

AI can deliver anything from video subtitles to digital transformation. But it can also have unintended consequences: making assessment less authentic, preventing students learning essential academic searching and evaluation skills, reducing opportunities for human and community contact and professional judgement, or introducing dataflows that have little to do with education or pastoral support. As benchmarks, compare the idea – primarily with educators and those involved in the affected processes – against relevant foresight documents such as Jisc's **Future of Assessment report** (jisc.ac.uk/reports/the-future-of-assessment), the institution's mission, Codes and Charters that the institution subscribes to¹, even the UN's Human Rights Declaration's statement of the purpose of education: "the full development of the human personality and to the strengthening of respect for human rights and fundamental freedoms". Consequences that are not clearly aligned with these are more likely to lead to ethical challenges.

¹ Eg StudentMinds Mental Health Charter (studentminds.org.uk/charter.html) or StandAlone's Charter for Estranged People (standalone.org.uk)

Step 2: Does using AI fit our institution's purpose and culture?

If the idea still seems plausible and fits institutional objectives, the next group looks at issues in using AI to implement it.

What level of ethical complexity can we accept?

Laws are starting to identify specific applications of AI as "high-risk" or, in a few cases, prohibited. For example the draft EU legal framework for AI (https://ji.sc/eu_ai_framework) considers processes that "may determine the access to education and professional course of someone's life (eg scoring of exams)" as high-risk, in particular of "perpetuat[ing] historical patterns of discrimination"; systems that "manipulate human behaviour to circumvent users' free will ... and systems that allow 'social scoring' by governments" are prohibited. Courts and Regulators have raised concerns about **face recognition and other forms of biometric identification** (<https://ji.sc/biometricid>), **automated decision-making** (https://ji.sc/auto_decision), and **potentially discriminatory data sources** (https://ji.sc/ai_data_sources). Types of data or processing identified as "special category" by the General Data Protection Regulation – racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, biometric, health and sexual – should also be considered high-risk. **Board** and **management** should consult with **stakeholders** (e.g. **academics, professional services, students**) before deciding whether they are ready for the legal, ethical and communications challenges of using AI in these ways.

Does using AI in this way fit our local purpose, community and culture?

Local Authorities have found that acceptable uses of AI are likely to vary between places, based on existing

circumstances, relationships and experiences. Extending this idea of "Place-based AI Ethics", from the PolicyConnect/ All-Party Parliamentary Group report "**Our Place, Our Data**" (https://ji.sc/ourplace_ourdata), we might ask **students, academics, tutors and professional services**: Would this contribute to the place you want to live, work and study? Does it involve the right mix of technological, cultural and community change? Does it retain, even enhance, important human contact? If it changes how you work, is that positive, or at least manageable? Are you confident we can make it work? Just because something is acceptable or even desired in one institution it may not receive the same response in others. Low-stakes pilots can help explore these questions.

Is AI a less intrusive way to do it?

Sometimes AI is less intrusive or risky than humans performing the same process. Students identify "study nudges" as something they would not want shared with their lecturer or teacher, for example. This strongly applies to behavioural data that wouldn't normally be available to any human, such as sleep times or patterns. Simple wellbeing reminders to **students** or **staff** to take a screen-break or not work at 3am might be more acceptable from an unmediated app, with humans only being informed if an unhealthy pattern emerges. But even data gathering or reuse that is acceptable in itself can contribute to a perception of gradually increasing surveillance. Check regularly with **stakeholders** that new and existing services are not approaching this point.

Step 3: Are we ready to do this?

While some uses of AI in education are standalone systems, others need incorporating into existing processes, from facilities management to libraries to teaching.

Institutional readiness will be critical to success. Two questions are well discussed in the **Ethical Framework for AI in Education** (https://ji.sc/ethical_ai_education).

Is the institution ready?

Successful uses of AI create a partnership between humans and machines, each doing the things they do best and with clear interfaces between them. This may involve significant change and place requirements on both sides. Do users of AI (**staff** or **students**) have the appropriate training, skills, resources and trust: being confident to rely on the AI when appropriate, but using human expertise to judge when it is not? Do the AI and supporting systems help by giving clear warnings of its limits and providing a route for corrective feedback when it makes errors or exhibits bias? Are there institutional

processes for when things go wrong? If the AI needs data or interfaces, do we agree those are appropriate (for example in meaning, quality and durability), and are they in the form that **implementors** require? Will staff and students need greater digital literacy and capability and, if so, will training or equipment be provided to avoid increasing existing divides?

Is the supply chain ready?

Off-the-shelf products and contracts may suit some situations, others may need more specialised support. If that support, documentation or functionality is not yet available, it's better to postpone the idea until either the market, or the institution's experience, has developed. Existing **supplier** relationships and **procurement** processes can provide a useful starting point for these discussions.



Step 4: Does using AI raise specific issues?

Some uses, and some organisational/technical contexts, may impose particular demands. Using AI that cannot meet those demands is likely to be high-risk. Always remember that “Artificial Intelligence” isn’t a different form of “Human Intelligence”, it is something else entirely. AI doesn’t “know” or “understand”: it calculates. The metaphor often misleads². Four areas to think about:

“Control”

Few, if any, instances of AI can operate independently for ever. Even the AI driving highly-autonomous oceanic and inter-planetary robots will need human help in an unforeseen situation, or if its environment changes so as to affect its original purpose. Before implementing AI, think when, and to what extent, humans need to be able to take over: does the context require periodic checks and tweaks, on-demand support or review, corrections to individual actions, or a complete change of approach (see **Learning** on the following page)? Can the AI itself ask a human to step in: flagging when it approaches the edge of its experience, for example? Control requires technical indicators and levers, and institutional processes to act on them, so needs discussion between **users, implementers** and **process owners**.

“Explanation”

For some uses of AI humans, and sometimes society, need explanations: whether “why did you use AI for that?”, “what goals did you set it?”, “is it fair?” (which, itself, needs defining), or “why did it do that to me?”. Explanations can highlight unconscious or learned bias (see on the following page). **Students, tutors, accreditors, stakeholders/funders** and others may expect different kinds of explanation; **communications experts** can help match explanations to their audience³. You may need to **change technology** (https://ji.sc/change_tech): when improving a process an approximate, but explained, answer is more use than a precision black-box. More widely: are explanations needed beforehand (eg explaining design choices) or after (eg validating that results were as intended), are they needed at population/cohort level (eg is the spread of outcomes fair?) or individual (what data error or behaviour do I need to change to get a different result?). Remember that some explanations seek to understand the institution or supplier and its processes and motivations, not a specific technology or decision.

² https://ji.sc/euoparl_metaphors

³ Eg <https://automated-decisions.tumblr.com>

Explanations (examples)	Before	After
Cohort	Will it allocate resources fairly? (for some value of fair)	Did it deliver the spread of outcomes we expected?
Individual	Will it disadvantage specific personas?	Why did it do that to me? What do I need to change?

“Bias”

AI can encode or amplify human bias. This is rarely intentional, but can result from **unrepresentative** (https://ji.sc/face_recognition) or **biased training or input data** (https://ji.sc/biased_data), or simply from assumptions about behaviour or data that **don't always apply** (https://ji.sc/ai_identifiers). For most applications it's unwanted, for some it's illegal (**professional services** can advise on these). But AI can also highlight bias and help us fix it. Before deploying AI, define what bias would look like in that context, design to avoid it, but also plan how to detect it and respond, including by reverting to a simpler algorithm or stopping using AI at all. **Educators** and **minorities** (including those with disabilities and widening access students) will often be most aware of bias, so their views, in particular, should be sought.

“Learning”

Every AI must “learn”, or be “taught”, about its environment, whether the statistical relationship between study hours and assessment outcomes or adapting to its interactions with humans. Depending on context, we may want to learn like a spreadsheet, a parrot, or a toddler!

You can prescribe the **inputs** AI learns from, the **methods** it learns with, and the range of **outputs** it can produce. Different choices can produce completely predictable or entirely unpredictable behaviour, so must match the requirements of the context. Unpredictable may be good for developing new insights, but disastrous if it reproduces biased or deliberately offensive human behaviour:

predictable may be desirable, even essential, if the goal is to reproduce past outcomes.

To show how this works, consider personalised reading recommendations to students:

- The most constrained version might use a student's previous modules (or school exam board) to suggest texts chosen by the tutor to fill in likely knowledge gaps
- Relaxing control of **inputs**, and perhaps **method**, an AI might analyse the student's work to identify areas of difficulty or misunderstanding
- Relaxing control of **outputs**, we might let students make recommendations (possibly inaccurate or malicious) to one another

12 | Step 4: Does using AI raise specific issues?

More generally, analysing some different learning approaches:

Type	Target	Prescribed	Risks include
Statistical	Find significant relationships between numerical inputs	Inputs: yes, raw data Method: yes Outputs: yes, model parameters	Data sources may be incomplete or biased (eg correlation between gender and job (https://jisc/paygap)). Relationships may be coincidental rather than causative. Humans may misinterpret/misuse results.
Curated	Find model to reproduce human classification of objects (eg tagged images)	Inputs: yes, curated training set Method: no Outputs: yes, labels	Incomplete/biased training set. AI may find unexpected ways to reproduce classification (eg COVID severity risk assessed based on whether an image was taken on a mobile (care home) or fixed (hospital) scanner).
Self-directed	Find algorithm that achieves a specified goal	Inputs: maybe Method: no Outputs: no	AI may find unexpected approach. Feedback risk if algorithm can influence data collection (eg predictive policing). Risk of side effects unless these are included in goal.
Experiential	Reproduce human behaviour observed during interactions	Inputs: no Method: no Outputs: no	Humans may teach it undesirable behaviour (eg Tay) (https://en.wikipedia.org/wiki/Tay_(bot))

Be particularly careful with approaches and contexts involving feedback loops. AI that can influence its own inputs – for example by sending humans to collect or generate more data – creates digital confirmation bias: diverting police officers into areas with high arrest rates is likely to produce even more arrests. AI that extends its range of outputs using what it observes or provokes can amplify undesirable features of its environment.

Finally, if unwanted behaviour does emerge, consider where that was learned from. If we can make the environment a better example, a short-term AI harm might result in a longer-term social gain.

Discuss these risks, and the appropriate level of prescription, with those familiar with, and affected by, the systems and environments from which the AI will learn, including **technologists, educators, minorities** (including those with **disabilities**) and **information security** (who will often know about problems elsewhere). Precautions against learning risks are similar to those for bias: good design, quick detection and correction of unwanted tendencies.

Finally: the details

Once an idea has made it this far, consider the details of law, technology and ethics. Discussions along the pathway, and information gained, should make these conversations much simpler.

Law

Many uses of AI will involve personal data, making the General Data Protection Regulation (GDPR) a legal requirement. Even those that don't can learn from the GDPR's Principles – lawfulness, fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality; and accountability – and its concerns with profiling and automated decision-making. Clarity about data use is good for AI and for acceptability. Data minimisation and storage limitation can seem a challenge, but given a clear purpose and lawful basis they actually benefit AI. A carefully-chosen set of input data is less likely to produce bias, discrimination and unforeseen effects than a “more-is-more” approach. The Information Commissioner has guidance on specific **Data Protection issues when using AI** (https://ji.sc/data_protection_ai2). Some AI applications also involve discrimination, equality and accessibility law; high-risk AI may affect Human Rights. Both UK and EU are working on AI laws.

Technology

The choice of AI technology – indeed, even whether and where to use AI – will often be coupled to these legal requirements. Considering the GDPR Principles: technology clearly contributes to integrity and confidentiality, but it can also help fairness and accuracy. Technologies such as pseudonyms and summary statistics make a significant difference to data minimisation and storage limitation, even to purpose limitation if raw, reusable, data can be transformed to limit possible (mis-)uses. Work with these principles: don't claim they are irrelevant because “technology makes data anonymous”. It rarely does. Don't underestimate the effort needed to get existing information into the form and quality needed by a live AI system.

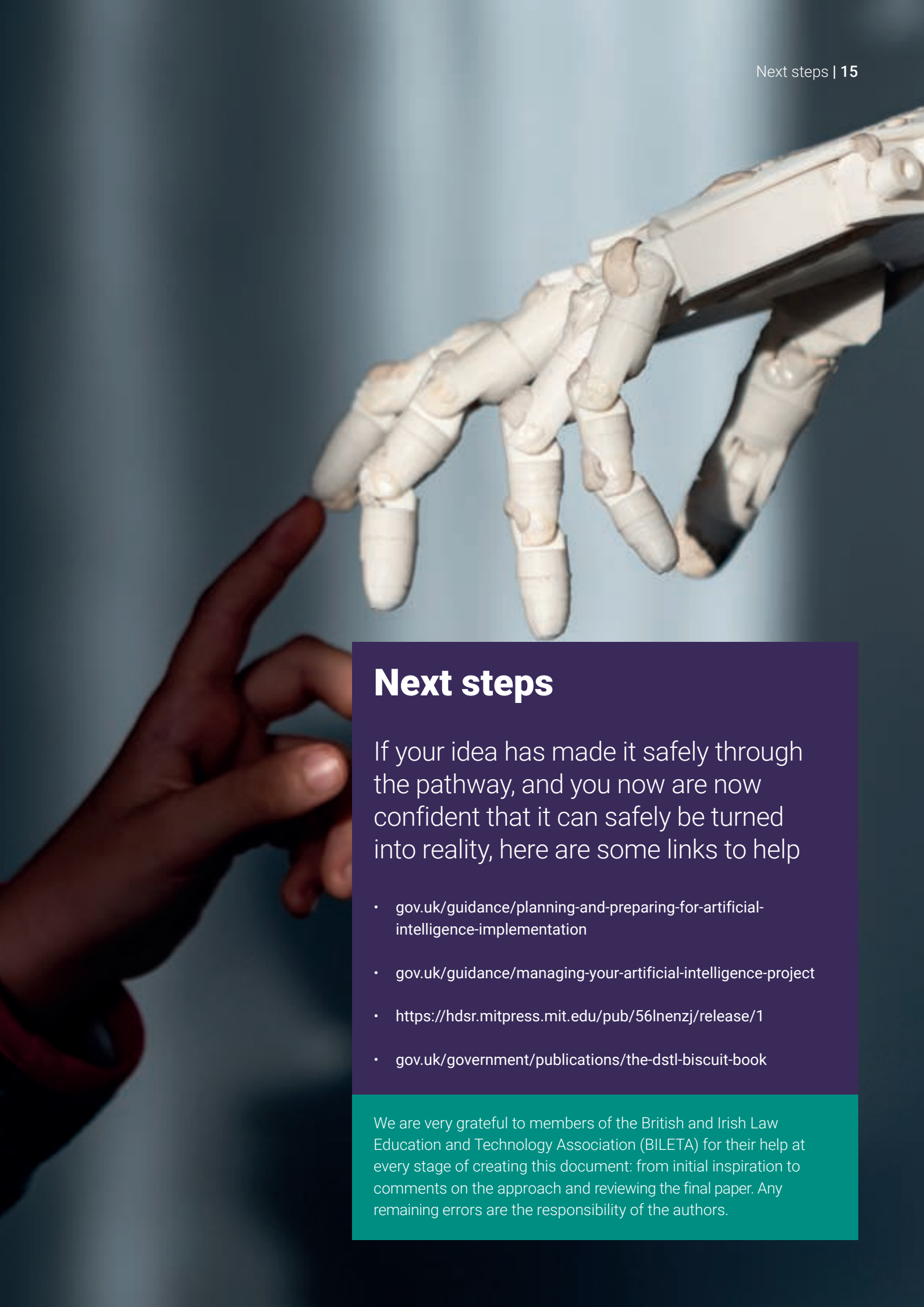
Ethics

Many AI Ethics codes have been written, but most share core principles – accountability, transparency, fairness, robustness/data quality, privacy, prevention of harm, respect for autonomy, etc. – that you'll already have discussed along the pathway (eg impact on groups, or the educational context) or in considering the GDPR and other legislation. High-risk applications of technologies and data should have been identified, and perhaps eliminated. Codes such as the **EU High-level Experts Group (on AI)** (https://ji.sc/EU_HLEG) or the **UK Government (on data)** (https://ji.sc/gov_ethics_framework) should hold few surprises. Hard ethical questions may remain (often, whether AI is used or not) where situations create conflicts between ethical principles, or between fundamental rights and freedoms. Here it is worth talking to professional **ethicists**, who can help you explore the trade-offs.

Summary of pathway

The pathway is not meant to involve a full risk assessment, though it may help you decide whether one of those is required. Its discussions may, however, identify risks, mitigations and benefits that are worth noting for later.

	Risks	Mitigations/benefits
1. Does this proposal fit our institution's objectives?		
Why are we doing this?		
Will it work?		
Does it advance our institution's mission?		
2. Does using AI fit our institution's mission and culture?		
What level of ethical complexity can we accept?		
Does using AI in this way fit our local mission, community and culture?		
Is AI a less intrusive way to do it?		
3. Are we ready to do it?		
Is the institution ready?		
Is the supplier ready?		
4. Does using AI raise specific issues?		
Control		
Explanation		
Bias		
Learning		
The details		
Law	Consider a DPIA to balance benefits and risks to individuals	
Technology	Consider how technology choices can affect your proposal's risks and benefits	
Ethics	Consider an ethical AI framework to identify and resolve ethical dilemmas	




Next steps

If your idea has made it safely through the pathway, and you now are now confident that it can safely be turned into reality, here are some links to help

- [gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation](https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation)
- [gov.uk/guidance/managing-your-artificial-intelligence-project](https://www.gov.uk/guidance/managing-your-artificial-intelligence-project)
- <https://hdr.mitpress.mit.edu/pub/56lnenzj/release/1>
- [gov.uk/government/publications/the-dstl-biscuit-book](https://www.gov.uk/government/publications/the-dstl-biscuit-book)

We are very grateful to members of the British and Irish Law Education and Technology Association (BILETA) for their help at every stage of creating this document: from initial inspiration to comments on the approach and reviewing the final paper. Any remaining errors are the responsibility of the authors.

Jisc
4 Portwall Lane,
Bristol BS1 6NB
0300 300 2212

help@jisc.ac.uk
jisc.ac.uk
 [@Jisc](https://twitter.com/Jisc)