Artificial Intelligence and Data Protection:
Delivering Sustainable AI Accountability in Practice

Second Report:
Hard Issues and Practical Solutions

January 17, 2020

Foreword

Dear Colleagues:

It is with great pleasure that I present to you the second report from the Centre for Information Policy Leadership's (CIPL) project on **Artificial Intelligence and Data Protection**.

Artificial Intelligence is transforming the way we live and interact in modern civilization and will continue to have a huge impact on the digital economy, particularly in the coming years as organizations, governments and wider society realize its full potential and reap its many benefits.

We have also seen an unprecedented transformation in the field of privacy and data protection, specifically towards the end of the last decade with the several jurisdictions overhauling and updating their privacy regimes for the new digital era and many others implementing comprehensive privacy laws and regulations for the first time. This dynamic legal and policy shift is set to continue and we will no doubt continue to see numerous global developments as new standards for protecting the privacy of individuals are set throughout our world.

Light speed progress in both the domains of artificial intelligence and data protection have created new issues, new questions and, sometimes, even tensions between these fields. In October 2018, CIPL outlined many of these key tensions through its first report in this project. In doing so, CIPL sought to start a global conversation around how we might be able to address the challenges that AI technology presents to data protection while still enabling innovation and advancement in this rapidly growing area.

Over the past year, CIPL has engaged in such discussion with organizations from different industry sectors, data protection regulators in North America, Europe and Asia, law and policy makers, academics and other key stakeholders. Through many roundtables, workshops, side meetings and conference calls, CIPL has identified a plethora of practical methods and measures to address several existing tensions, and that users of AI technology can implement today to ensure that they are not only in compliance with data protection requirements but that they are truly delivering sustainable AI accountability in practice.

I hope you find the information in this report useful and that it serves your organization as you navigate issues of data protection and its interaction with AI technologies.

I would like to take this opportunity to thank CIPL members as well as all the participants in our numerous events over the past year for their contribution to the discussions and in making this second report come together. We look forward to continuing our collaboration together as we move to the next phase of this project, which will look more deeply at AI governance mechanisms, layered approaches to regulating AI and specific technologies, including facial recognition.

For more information about CIPL's project on Artificial Intelligence and Data Protection, please visit https://www.informationpolicycentre.com/ai-project.html or reach out to me directly at bbellamy@HuntonAK.com

Sincerely,

Bojana Bellamy
President

Table of Contents

I. Executive Summary 4

II. Understanding the Issues 5

A. Fairness 5

B. Transparency 9

C. Purpose Specification and Use Limitations 12

D. Data Minimization 13

III. Solutions Ahead 15

A. The Need for Technology-Neutral Solutions 15

B. The Importance of Process 16

C. A Risk-based Approach to AI 17

D. The Need for Data Stewardship and Organizational Accountability 18

E. Focusing on Meaningful Roles for Humans 19

F. Wide Range of Available Tools 19

1. CIPL Accountability Wheel 19

2. AI Data Protection Impact Assessments (AI DPIAs) 20

3. Data Review Boards (DRBs) 21

4. Avenues of Redress 22

IV. Conclusion 23

I. Executive Summary

In October 2018, CIPL published the first report of its *Project on Artificial Intelligence and Data Protection: Delivering Sustainable AI Accountability in Practice*. The report detailed the widespread use, capabilities, and remarkable potential of AI applications and examined some of the tensions that exist between AI technologies and traditional data protection principles. The report concluded that ensuring the protection of personal data will "require forward-thinking practices by companies and reasonable interpretation of existing laws by regulators, if individuals are to be protected effectively and society is to enjoy the benefits of advanced AI tools."[1]

Following publication of the first report, CIPL has taken a deeper dive into examining some of the hardest challenges surrounding AI, including by hosting roundtables and workshops around the world with regulators, law and policy makers, industry leaders, and academics to identify tools and emerging best practices for addressing these key issues.[2] This approach has not attempted to provide a comprehensive analysis of every issue. Rather, we have focused on particularly critical issues for ensuring the responsible use of AI in the context of data protection laws that were often enacted prior to the explosion in AI technologies. These issues include fairness, transparency, purpose specification and use limitations, and data minimization.

In these workshops and other conversations with AI and data protection experts and regulators around the world, six messages have emerged with remarkable consistency:

1. The proliferation of AI tools, their expanding impact on individuals and societies, and their reliance on large volumes of granular, often personal data, require **effective data protection** by both the private sector and governments.

2. There is sufficient scope in current data protection measures to provide much of that protection. However, achieving that requires **creativity, flexibility, agility, cooperation, and continued vigilance** from organizations, regulators and policymakers, as AI technologies and applications, as well as public perceptions and our understanding of the risks, evolve.

3. In some areas, including some of the issues we address below, innovations in **governance and regulatory approaches** likely will be needed to ensure that individuals and communities can enjoy the full potential of AI without compromising the protection of personal data or other fundamental rights. Fortunately, AI may help facilitate new approaches, and it is important to ensure that those approaches are consistent with, and not duplicative of, existing regulations.

4. The process of **creating, innovating, and collaborating across multiple disciplines and teams** is critical to achieving the responsible use of AI and the protection of personal data, and will help equip us to anticipate and respond to other data protection challenges in the future.

5. We must be **reasonable in our expectations of AI**, especially at the start. We, as humans, are rarely consistently rational, unbiased, or even capable of explaining why we reach the decisions we do. In the long run, AI is far more likely than humans to achieve the goals of rationality, consistency, and fairness, and we should aspire for AI to do so. However, if we insist and require, from the beginning, that AI meet standards that human behavior cannot, we run the risk of restricting the development of new tools with enormous potential for individuals and society.

6. Finally, many **organizations and leaders in AI technology are proactively starting to address the risks, challenges, and tensions** to deliver accountable AI and in compliance with data privacy laws and societal expectations. These emerging best practices are starting to take shape in the form of more coherent and comprehensive accountable AI frameworks, including on the basis of the CIPL Accountability Wheel (see Part III.F. below).

This second report provides an overview of what we learned about the AI challenges in the context of data protection; concrete approaches to mitigating them; and some key examples of creative approaches and tools that can be deployed today to foster a better future in which human-centric AI, privacy, and both the protection and productive use of personal data prosper.

In the next section, we discuss four significant data protection challenges presented by AI and provide examples of tools for managing them. In the final section, we describe broader cross-cutting themes that have emerged from our research, as well as best practices, controls, and tools that are helping to resolve or mitigate these challenges. We also detail how the CIPL Accountability Wheel may be a useful framework for organizations and regulators to structure these best practices in a way that delivers trustworthy and accountable AI in practice.

II. Understanding the Issues

Considerable advances in AI have created or exacerbated challenges to existing data protection principles, some of which are proving particularly difficult and important for organizations and regulators to address. This section will analyze four of the most challenging issues that regulators, industry leaders, AI engineers, academics, and civil society are grappling with as they consider AI and data protection. We will also offer some of the interpretations and solutions being used to strike an appropriate balance between the proliferation of AI applications and the protection of personal data.

A. *Fairness*

Fair processing is a fundamental data protection principle and a requirement of the EU General Data Protection Regulation (GDPR) and other data protection laws.[3] Yet, despite its importance, the principle of fair processing has not been authoritatively or consistently defined. Defining fairness has been an ongoing challenge both in the context of AI and elsewhere in privacy and data protection. The longstanding test for what is an "unfair" business practice employed by the US Federal Trade Commission is whether the practice causes a substantial injury that is not outweighed by any countervailing benefits to consumers or competition that the practice produces and that causes an injury that consumers themselves could not reasonably have avoided.[4]

The EU appears to regard "fairness" as a much broader concept. The 23 official languages of the European Union into which the GDPR principles have been translated suggest a wide range of meanings, including good faith, honesty, propriety, goodness, justice, righteousness, equity, loyalty, trustworthiness, fidelity, objectivity, due process, fair play, integrity, reliability, dependability, uprightness, correctness, virtuousness, justice, and devotion (see Appendix A). The European Data Protection Board's recent draft guidelines on *Data Protection by Design and by Default* attempt to advance the interpretation of fair processing. They define fairness as requiring that personal data shall not be processed in a way that is "detrimental, discriminatory, unexpected or misleading to the data subject," and further outline twelve elements of a fairness assessment: autonomy, interaction, expectation, non-discrimination, non-exploitation, consumer choice, power balance, rights and freedoms, ethical, truthful, human intervention,

and fair algorithms.[5] These may indeed all be desirable attributes.[6] However, as a foundational principle for data protection and a legal requirement under the GDPR, some broader deliberations among stakeholders on the concept of fair processing (or fairness and unfairness) would be desirable for both regulators and regulated organizations.

In practice, fairness appears to be an amorphous concept that is **subjective, contextual, and influenced by a variety of social, cultural, and legal factors**. The same data used in different contexts may raise entirely different reactions to fairness questions. For example, if universities use prospective student data to train an algorithm that tailors advertising to "non-traditional prospects" such as first-generation university students, the assessment of fairness may be different than if the same data is used to identify students most able to pay for university and direct advertising toward those more well-resourced populations.[7] The contextual nature of fairness creates significant challenges for regulators charged with interpreting and enforcing the law, for organizations charged with implementing it, and for individuals whose rights are supposed to be protected by it.

The difficulty and importance of defining and ensuring **fairness is only magnified in AI contexts**. This is true because of the scale, speed, and impact of AI; the complexity of AI algorithms; the variety and sometimes uncertain provenance of input data; the unpredictability or sometimes unexpected outcomes from certain AI algorithms; a frequent lack of direct interaction with individuals; and less well-formed or defined expectations of the average individual. These characteristics of AI often exacerbate the challenges of fairness not having a clear and consistent meaning, and of that meaning depending, at least in part, on context and other factors.

Another potential challenge of fairness is **the lack or invisibility of an individualized harm**. Unfair outcomes are often broad-based impacts to society as a whole, and even where individual harm occurs, it is not easily recognized at an individual level. This creates an imperative for more action by organizations and regulators to safeguard fairness. For example, at the current state of development, facial recognition technologies tend to be more accurate for lighter-skinned individuals, so deploying these technologies to a diverse population in high-risk situations could create unfairness. The largest US provider of police body cameras has decided to stop using facial recognition technology on police vests because it believes the likely inaccuracy, and systematically discriminatory impact, in such a high-stakes setting is unethical.[8] Similarly, the UK Information Commissioner's Office (UK ICO) has advocated for police forces to "slow down," to consider the impacts of live facial recognition, and to take steps to eliminate algorithmic bias prior to deployment.[9]

As these examples suggest, fairness should be addressed from two dimensions: fair process (meaning processes that take into account the impact on individuals' interests) and fair outcome (meaning the appropriate distribution of benefits). Both dimensions need to be addressed if we are to maximize the value of data and its applications for all those with an interest in it. The principles and values existing in most data protection regulations remain relevant, but now there is a need for more dialogue and an exchange of views among stakeholders to create practical ways of fulfilling these principles.

Fortunately, both regulators and organizations are striving to facilitate **progress towards greater fairness in the development and deployment of AI technology**.

a) The European High-Level Expert Group (HLEG) on AI published *Ethics Guidelines for Trustworthy AI*, encouraging organizations to "ensure an adequate working definition of 'fairness'" to apply in AI systems design.[10] The Guidelines provide **questions for companies to consider when creating**

policies to promote fairness, but ultimately allow organizations to develop their own definition and approach to fairness as well as the processes and mechanisms to achieve it.

b) Another approach from regulators could be to **define outcomes of fairness or considerations for fairness**. For example, the UK Financial Conduct Authority (UK FCA) requires all firms under its authority to treat customers fairly, and toward that end, it has defined six outcomes of fairness that organizations should create policies and procedures to achieve:

Outcome 1: Consumers can be confident that they are dealing with firms where the fair treatment of customers is central to the corporate culture.

Outcome 2: Products and services marketed and sold in the retail market are designed to meet the needs of identified consumer groups and are targeted accordingly.

Outcome 3: Consumers are provided with clear information and are kept appropriately informed before, during and after the point of sale.

Outcome 4: Where consumers receive advice, the advice is suitable and takes account of their circumstances.

Outcome 5: Consumers are provided with products that perform as firms have led them to expect, and the associated service is both of an acceptable standard and as they have been led to expect.

Outcome 6: Consumers do not face unreasonable post-sale barriers imposed by firms to change product, switch provider, submit a claim or make a complaint.[11]

These fairness outcomes essentially assess and balance risk for the industry, requiring organizations to build processes to achieve them.

These regulatory approaches of the European HLEG on AI and the UK FCA allow organizations to innovate on how they define and safeguard fairness while also providing some guidance on what it means to be fair.

A key tool for assessing and achieving greater fairness is **the use of a risk- or harm-based method to guiding decisions**. This standard focuses on the impact of uses of data and potential for harm to individuals rather than the expectations of a hypothetical reasonable person. Certain types of decisions bring different levels of risk or potential for harm, such as the difference between recommending a traffic route for logistics versus making a healthcare decision. To determine the level of risk, an organization may account for the population impacted, the individual impact, and the probability or potential for harm. As the potential for harm or discriminatory impacts for individuals increases, so should the controls and checks put in place by an organization to limit negative consequences to fairness. This is also the very essence of the approach under CIPL's Accountability Wheel framework (discussed in detail in Part III.F. below): organizations must calibrate their privacy program and specific controls based on the outcomes of the risk assessments they conduct. The higher the risk, the more they must do by way of oversight, policies, procedures, training, transparency, and verification.

However an organization defines and assesses fairness, it is important to note that **fairness is not absolute and may require continual and iterative reassessment**. Throughout conversations with regulators and industry leaders, a common view focused on fairness as existing on a spectrum rather than being a binary concept. Participants identified a number of measures, described in greater detail below, that could be used to make the development and deployment of AI "more fair," but none of these measures will provide a silver bullet to achieving fairness. Rather, ensuring fairness is a continuous process. What that process looks like may depend on the context, the culture, or the organization, but the development of processes and monitoring of outcomes will help organizations move along the spectrum toward fairness regardless of the specific understanding and definition of fairness that is being applied.

Organizations are starting to develop a number of **technical and procedural tools and frameworks** to help ensure fairness in AI applications.

- In particular, organizations are developing **techniques** to identify and address the risk of algorithmic bias, as they rightly perceive bias to be one of the key indicators of unfairness. One example is counterfactual fairness testing, which checks for fairness in outcomes by determining whether the same result is achieved when a specific variable, such as race or gender, changes.[12] In fact, as identified in CIPL's first AI report[13] and in numerous discussions with AI engineers, it is absolutely necessary for organizations to process and retain sensitive data categories, such as ethnicity or gender, to prevent bias in the model, or to be able to test the model, monitor and fine-tune it at a later stage and, thus, ensure fairness.[14] Google, for example, has developed algorithmic fairness techniques to "surface bias, analyze data sets, and test and understand complex models in order to help make AI systems more fair," including Facets, the What-If Tool, Model and Data Cards, and training with algorithmic fairness constraints. These and other tools are described in greater detail in Google's 2019 report, *Perspectives on Issues in AI Governance*.[15] Accenture has developed a fairness tool to "identify and remove any coordinated influence that may lead to an unfair outcome,"[16] which it uses both internally with respect to its own AI projects, as well as externally, on client projects involving the deployment of AI applications to help clients address the fairness standard. IBM has also created several tools to address issues of ethics and fairness in AI, including AI Fairness 360, "a comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias,"[17] as well as IBM Watson OpenScale, a tool for tracking and measuring outcomes of AI to help intelligently detect and correct bias as well as explain AI decisions.[18]

- Equally important as these technical tools are the variety of **procedural and accountability mechanisms** to ensure fairness. Organizations can create internal governance structures and accountability frameworks, and then utilize tools such as AI data protection impact assessments (AI DPIAs) or data review boards (DRBs) to implement AI accountability (discussed in Part III of this report). These mechanisms are particularly useful in the development phase of AI applications, but also in the review and monitoring phases. Of course, providing transparency and mechanisms for redress will be essential to ensuring fairness throughout the deployment of AI technologies. All of this exemplifies the point that **fairness has to be ensured throughout the lifecycle** of an AI application – from evaluation of the AI use case and input data, algorithmic modeling, development, and training to deployment, ongoing monitoring, verification, and oversight.

- Furthermore, many consider that transparency, explainability, and redress are intrinsically linked to the assessment of fairness in an AI application. In other words, providing for user-centric and

meaningful transparency and explainability of the AI decision-making process and enabling redress to individuals is likely to increase the chances that a specific data processing in AI is fair.

One key challenge when conducting a fairness assessment is whether an organization or regulators can or should **consider issues outside of data protection**. Should organizations or regulators look beyond data protection issues to assess broader potential impacts of AI, for example, on the future of work or the competitiveness of firms? Or should they consider the impacts on other human rights, beyond data protection, when considering the fairness of the AI application?

Often, there are other bodies of law tasked with assessing these concerns, and individual organizations, much less data protection offices within organizations, are often poorly equipped to address broader impacts on societies at large. After all, even trade ministries and other government policymakers, with their broad missions and resources, often have difficulty predicting the social and economic impact of new technologies.

However, the GDPR seems to suggest that assessing fairness requires consideration of issues outside of data protection. For example, data protection impact assessments (DPIAs), discussed further below, are required to evaluate processing "likely to result in a high risk to the rights and freedoms of natural persons,"[19] and this is not limited to evaluating data protection or privacy rights. Furthermore, the 41st International Conference of Data Protection & Privacy Commissioners (ICDPPC) called upon all organizations to "assess risks to privacy, equality, fairness, and freedom before using artificial intelligence."[20]

Given these considerations, it is clear that an organization looking to holistically assess an AI application or new data use might want to extend the scope of their lens to include such broader issues. This may also be expected as part of the organization's broader corporate social responsibility, of which digital responsibility is a subset. A small number of organizations are even working to develop a broader human rights impact assessment that they will conduct when developing and deploying AI technology. These tools are still very much in the early stages of development, but indicate the direction of travel for some accountable organizations.

Finally, while defining and implementing fairness is a challenge, it is also an **opportunity**. As Google recently noted, "[i]f well implemented, an algorithmic approach [to fairness] can help boost the consistency of decision-making, especially compared to the alternative of individuals judging according to their own internal (and thus likely varying) definitions of fairness."[21] AI can ultimately help facilitate the goals of fairness – either by helping to illuminate and mitigate historical biases or providing more consistent and rational decision-making. Achieving this goal will require that organizations define fairness and develop tools and procedures to uphold and ensure their definition throughout the process of AI development and deployment.

B. *Transparency*

Transparency, like fairness, is a concern exacerbated by AI, but it is also a potential solution for many of the fears around AI technologies. Transparency regarding AI requires "organisations to provide individuals the specifics of data processing, including the logic behind any automated decision-making that has legal effect or a similarly significant impact on individuals."[22] The goals of transparency are to inform individuals about how their data is used to make decisions, hold organizations accountable for their policies and procedures concerning AI, help detect and correct bias, and generally foster trust in the use and

proliferation of AI. The tools regulators and organizations rely on to facilitate transparency should be developed to serve these goals.

Transparency has been a difficult challenge in AI, as it is often unclear what we mean by transparency. As a starting point, it can be helpful to consider the human decision-making alternative. Often, humans are unable to consistently explain their preferences for one option over another, and there are a number of situations where decisions are not completed in a transparent manner, such as loan or credit approvals or hiring decisions. While we may be able to subsequently ask for an explanation, this explanation may not be logical, and almost certainly not technical or mathematical. Considering approaches to transparency in an offline world can be illustrative of what level and type of transparency to strive for when building AI systems.

The challenge of transparency in AI is made more difficult due to the complexity and changing nature of AI algorithms. One of AI's strengths is spotting complex patterns that had previously been missed, but such complexity by nature is inherently hard to explain in terms that are easily understood by humans. Advances in research have led to tools that can help developers better understand how their AI models work, but this requires investing the time and energy to interrogate models, which may not always be feasible. Furthermore, AI systems may be updated and re-trained using additional inputs, so decisions may not be easily repeatable. Because these systems are complex and often changing, providing information about the algorithm may not serve the goals of transparency. Not only is disclosing the code to individuals unlikely to be particularly useful for providing clarity about the decision, but algorithmic transparency could have the potentially harmful effects of disclosing trade secrets or helping individuals game the system.

Nonetheless, transparency is a legal obligation under the GDPR and other data protection laws[23] and a useful tool for building trust in AI. Hence, it is essential to build consensus on what transparency means in any given situation and to find ways to provide effective and meaningful transparency that achieves the goals mentioned above.

Some have argued for mandatory **disclosure of the fact that AI is being used** when an individual is interacting with a machine and for an explanation of why AI is being deployed and what is expected from its use. This could be helpful in some cases, but it will often be obvious, overly burdensome, or otherwise ineffective in building trust. Ultimately, it is not the technology that matters, but rather the fact that a non-human decision is having consequences on an individual in a way that he or she might not expect.

Transparency may differ depending on the audience it is geared toward – the individual or category of individuals impacted by the decision, the regulator, a business partner, or even for purposes of internal transparency to an oversight board or senior leaders. All of these different audiences imply different types and requirements of transparency that should be fulfilled appropriately. For example, a regulator may need to know more details about an AI use-case in the context of an investigation or audit – the model, the data sets, input and output, measures to ensure fairness and absence of bias, etc. On the other hand, for an individual, this type of information may be too much and "missing the forest for the trees." Equally, an organization developing AI technology to be used by another organization may be unable to provide transparency to data subjects directly, but it may need to provide additional transparency about the technical measures to ensure proper working of the model, bias avoidance, accuracy, documentation regarding tradeoffs, etc. Therefore, it may be hard to be categorical about the precise list of elements of transparency, as it very much depends on who the audience is and the specific purpose of transparency in a given context.

The level and method of transparency should ultimately be tied to the context and the purpose of AI applications. For example, the UK ICO's recent Project Expl*AI*n, a collaborative project between the ICO and the Alan Turing Institute to create practical guidance to assist organisations with explaining AI decisions to individuals, surveyed citizen juries and empirically demonstrated that individuals facing AI healthcare scenarios cared more about accuracy than transparency, while transparency expectations were heightened for the use of AI in job recruitment and criminal justice scenarios.[24] This suggests that transparency, and the tools used to achieve it, may differ based on what the AI application is used for, what the consequences are, and what rights individuals have going forward.

To illustrate these different considerations for transparency, consider the use of facial recognition technologies by airlines to check boarding passes or by customs officials to allow individuals into a country. The decision made by the AI in these cases is very significant, but transparency regarding the fact that AI is being used or about the code itself is unlikely to be of concern to the impacted individual. Instead, the concern is with how to contest or change the decision, so facilitating the goals of transparency will require a greater emphasis on speedy and effective avenues of redress. While human review is a requirement for certain impactful automated decisions under the GDPR,[25] developing efficient and visible avenues for such review – whether before or after a decision – will be an important part of transparency in AI contexts.

The level of transparency and the amount of human intervention needed may **vary depending on the risk posed to the individual by a decision or the visibility of the processing**. To conceptualize this point, it may be helpful to consider a four-box quadrant on impact and visibility (Figure 1). Some decisions require greater transparency due to their risk of harm, while others will require greater transparency due to their invisibility. Some may require little or no additional transparency, such as recommendations for restaurants or movies.

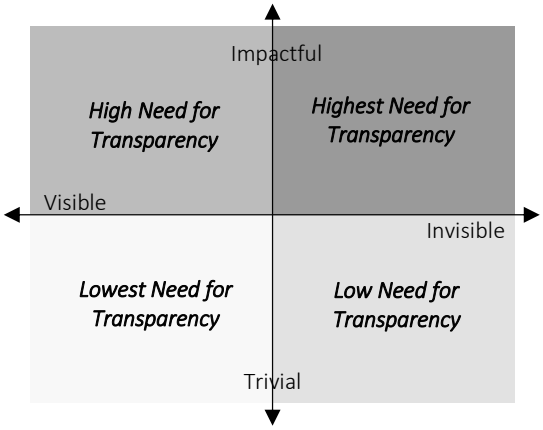

Figure 1. The figure to the left demonstrates the need for transparency based on the visibility of processing or use of technology and how impactful or how great the risk of harm is to an individual. Naturally, individuals and regulators alike are most concerned with processes that pose a greater risk of harm, such as consideration for a loan or insurance. However, transparency is also critical when there is no evidence of processing, such as with political marketing ads. When there is little to no risk of harm (or if the impact of that harm is trivial), there is less of a focus on transparency.

It is also clear that **transparency is a broader concept in the context of AI** – it includes explainability and understandability, as well as transparency concerning redress options and the ability to contest an AI decision. The EU HLEG on AI considered transparency to include elements of "traceability, explainability, and communication."[26] Traceability requires documenting data inputs and other "data sets and processes that yield the AI system's decision."[27] Creating codes of conduct or best practices regarding the collection, deployment, and use of data can help improve traceability. This can assist with ensuring fairness, conducting audits, and fostering explainability. As noted in the Model AI Framework created by the Singapore PDPC, "[a]n algorithm deployed in an AI solution is said to be explainable if how it functions and arrives at a particular prediction can be explained."[28] Tools such as using counterfactuals, factsheets

(which provide information or characteristics about AI services),[29] or Model Cards (short documents accompanying AI models that may describe the context in which models are to be used, evaluation procedures, and other relevant information)[30] can be useful for explaining decisions. Overall, traceability and explainability are ways of providing transparency about the AI outcome or data processes without providing transparency about the algorithm itself. These concepts promote the goals of transparency by increasing trust and accountability in decisions.

Finally, transparency also means the ability to **articulate benefits of a particular AI technology** and tangible benefits to individuals, as well as to broader society. In this way, organizations are able to provide educational value to individuals and drive greater trust and acceptance of these new applications.

While organizations can take steps toward providing more effective transparency regarding inputs and expectations, there may be circumstances where decisions cannot be explained to the degree necessary for regulators and individuals to be confident in AI decision-making. In this case, organizations should use other methods to foster this confidence and trust in AI. The most important tool to help with this will be providing visible avenues for redress through organizational accountability, which will be discussed in greater detail in Part III of this report.

C. *Purpose Specification and Use Limitations*

The principles of purpose specification and use limitation respectively require companies to specify the purpose for which they are processing data and then use data only for that or a compatible purpose. These principles have already been challenged by the prevalence of big data, but have recently been challenged by AI as well. Both companies and regulators must evaluate how to meaningfully apply purpose specification and use limitation in a way that serves the goals of these principles, while also allowing society to benefit from the capabilities of new technologies.

It is important to note that the principles of purpose specification and use limitation are not absolute. For example, the purpose limitation principle of the GDPR requires that personal data be "collected for specified, explicit and legitimate purposes*, and not further processed in a manner that is incompatible with those purposes*."[31] The OECD Privacy Guidelines, which undergird most modern data protection laws, contain similar language.[32] These principles are designed to limit unforeseen or invisible processing of data, so allowing for compatible processing serves the spirit of the principle while also allowing some flexibility.

The challenge presented by AI is due to its ability to sometimes discover unexpected correlations or draw unforeseen inferences from data sets. This may expose new uses or purposes for old data. For example, new computer vision technologies may be able to use old medical scans and charts to develop correlations and discover new value in old data. Use limitation and purpose specification, if interpreted narrowly, could stifle further research and preclude individuals and society from recognizing some of the potential benefits of AI.

Organizations and regulators must balance providing meaningful purpose specification and use limitation while also providing flexibility to react to new inferences from old or different data sets. Broad purpose or use specification statements provide little meaning for individuals and may ultimately degrade the effectiveness of these principles. The spirit of purpose specification requires that notice be precise, as "use for AI" alone would be neither specific nor precise enough to provide meaningful information to the individual.

Instead of allowing purposes to become so broad as to be meaningless, data protection authorities have interpreted purposes narrowly, which highlights the need to provide flexibility for allowing further processing. The GDPR, like the OECD Privacy Guidelines, explicitly permits further processing for new, "not incompatible" purposes.[33] The GDPR criteria of what is "not incompatible" are helpful in allowing for future uses.[34] Ultimately, further processing based on "compatibility" should be allowed for future uses that are consistent with, can co-exist with, and do not undermine or negate the original purpose. These uses must be backed by strong accountability-based safeguards, including benefit and risk assessments, to ensure that new uses do not expose the individual to unwarranted increased risks or adverse impacts. Indeed, the GDPR itself lists "the possible consequences of the intended further processing for data subjects"[35] as one consideration of the compatibility assessment. This is, in effect, a risk-based approach to determining what is "not incompatible." The higher the risk of adverse consequences, the less compatible the further processing is and vice versa.

A helpful distinction would allow **training an algorithm to serve as a separate and distinct purpose**.[36] The concept of a training phase is novel to AI, and data is often needed in greater amounts during the training phase than during deployment. In the training phase, where no individual decision-making occurs, the risk of harm to individuals by repurposing their data is lessened or eliminated entirely. As such, further processing in this phase should be deemed to be compatible with the original purpose. In addition, processing of personal data for the purpose of AI/model training could be a good example of processing based on the legitimate interest balancing test under the GDPR and other laws that have a similar provision.

Finally, the level of continued notice and the requirements necessary for further processing old data may be understood as a function of the risk of harm posed by that processing. "Data used in one context for one purpose or subject to one set of protections may be both beneficial and desirable, where the same data used in a different context or for another purpose or without appropriate protections may be both dangerous and undesirable."[37] Therefore, purpose specification and use limitation may be more effective if these principles rely less on evaluating the appropriateness of an intended use by the original terms and instead focus on the risk and impact of the new use. This focus on data impacts will be further explored in Part III of this report.

D. *Data Minimization*

Data minimization poses a similar paradox: a principle to limit collection and retention of individuals' personal data has the potential to stifle advancements in AI that could ultimately be beneficial to society. As mentioned above, AI has the capability of finding new and beneficial uses for old data, so it may be impractical to minimize data collection or retention.

While the intention and goals of the data minimization principle are still possible in our technological landscape, achieving these goals will require more creative solutions and flexible interpretations. The GDPR, for example, requires that: "Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[38] For AI, particularly at the development and training stages, what is *necessary* is a considerable amount of data, and having too little data can hinder the development of an algorithm. This is counterintuitive to the traditional understanding of data minimization, as having more data – in some cases – is necessary to reduce privacy risk. Thus, data minimization should not necessarily mean less data, but rather that data collected and retained is necessary in relation to the purposes of processing.

Throughout the CIPL roundtables examining this issue, participants noted that data can sometimes become less valuable as it ages – by losing either accuracy or relevancy. However, in certain sectors or for certain purposes, old data is invaluable, either to identify trends or train algorithms. In the financial industry, for example, old data can reveal patterns and identify trends that were unknown at the time of collection, and this can be particularly useful for fraud prevention. In the healthcare industry, old data can be useful in training algorithms to read scans or medical charts to detect patterns and learn more about disease prevention. Analyzing old data using new AI tools can create numerous opportunities and benefits, and in many of these cases, the future use of the data is agnostic regarding the identity of the individual.

Finding the appropriate balancing of interests between protecting individuals' data and providing avenues to reuse old data is a challenge for regulators and organizations. This challenge will require **flexible and reasonable interpretations of the data minimization principle**, often by distinguishing between the various contexts and purposes for storing data.

- Differentiating between data used for training AI versus deploying AI is helpful in this context. While other accountability tools will be necessary to govern the training phase, distinguishing between the training and deployment phase for purposes of data minimization could help balance innovation while fostering better data protection for individuals. By limiting data use in the deployment phase but providing more flexibility for data use in the training phase, organizations are managing the potential harm to individuals and thus upholding the original intention of the data minimization principle. This does not suggest that data is unnecessary in the deployment stage; it may in fact be critical for optimizing algorithmic performance and monitoring outcomes. However, the risk of harm to individuals is lessened during the training phase, so the standards for using data during that phase may be scrutinized less strictly. Another potential way to limit the use of personal data is to use synthetic data (i.e. a repository of data that is generated programmatically), when available and reasonably affordable, to train the AI model.

- Another reasonable interpretation of this principle would find it permissible to demonstrate data minimization with respect to an AI system by proactively articulating and documenting the need to collect and process data (whether it is old data or data not on its face strictly necessary to the purpose of the processing), and what is expected to be learned or accomplished by processing the data. This would be especially helpful for the training phase, although it could be useful for both training and deployment. Determining what is adequate, relevant, and necessary will be dependent on the context, but this proactive and continuous assessment will serve to demonstrate that the data to be collected is relevant and not excessive in relation to the purpose for processing.

- Throughout the lifecycle of an AI application, organizations can consider deploying a variety of tools to minimize risks to individuals. Technological tools to help with data minimization are still in an early stage of development and are often expensive for smaller organizations to deploy, but their continued exploration should be encouraged.[39] For example, in some cases, federated learning could enable AI algorithms to learn without data ever leaving a device and without the need to centralize large amounts of data in a single virtual location. Organizations may also consider the possibility of anonymizing or pseudonymizing data sets, although this may pose challenges of its own.[40] At the same time, while further research and development efforts are needed to ensure proper de-identification, a flexible interpretation of notions of anonymous or

pseudonymous data would go a long way to enable use of data for training of AI and to reduce the compliance risks for organizations.

III. Solutions Ahead

This second report has surveyed some of the most difficult tensions between AI and data protection and explored ways of mitigating those tensions. Throughout our conversations and roundtables examining these issues, six overarching themes have emerged. These cut across the issues already discussed and provide useful guides to developing, implementing, and assessing practical solutions for the responsible use of AI technology.

A. *The Need for Technology-Neutral Solutions*

Most of the data protection challenges identified in the AI context both predate AI and are posed by technologies other than AI. In short, those challenges are bigger and broader than AI, so it is important that the solutions are too. AI-specific solutions may not only be too narrow, but may also address a symptom without resolving the underlying problem. For example, the discomfort that may result from automated decision-making using AI is likely not the result of the technology itself, but rather the fact that a machine is making a significant decision that could negatively impact an individual, or even create legal effects for individuals. While AI may aggravate these issues, for example, by empowering machines to make more decisions affecting individuals, the problem to be addressed is not the AI, but rather the role of non-human decision-making, especially where these are significant decisions. The type of technology is nearly irrelevant; the impact of the decision made by that technology is the source of discomfort or distrust. Therefore, the solution should focus on the problem, not the technology.

As countries and regions consider regulation around AI,[41] it is important to understand that AI-specific legal structures or regulations could fail to resolve the underlying issue, while at the same time potentially deny society the benefits of properly implemented AI. This would also potentially deny society of the benefits from AI that does not involve automated decision-making. Additionally, any type of regulation that is not technology-neutral may overlap with or duplicate already existing (horizontal) regulations, which would be detrimental to legal certainty. AI-specific regulation runs the risk of creating harms without corresponding benefits, and where AI regulation is unavoidable, it should be developed thoughtfully and with enough time to allow a variety of stakeholders to identify, articulate, and implement key principles and best practices.[42] In the short term, however, serving the goals of enhancing data protection will instead require technology-neutral solutions and tools that can be applied across a variety of situations and contexts.

Technology-neutral solutions will help serve the goals of data protection in a more holistic way. As recently pointed out by the Platform for the Information Society Report, "[M]ost AI does not work independently: it is part of a product or service."[43] Technology-neutral tools will serve to improve the product, process, or service as a whole rather than one segment of a broader context. The UK FCA Outcomes for Fairness, discussed above, are one example of how technology-neutral tools can help facilitate overall responsible behavior by organizations. Organizations cannot be relieved of their responsibility by changing technologies; they must achieve the outcomes of fairness – and other data protection principles – irrespective of the technology or process used.

B. *The Importance of Process*

AI tools are being applied widely and take advantage of technologies that are developing at a rapid rate. Therefore, approaches towards solving the data protection challenges they may raise not only need to be technology-neutral, but also more focused on decision-making and remediation processes. What are the processes that an organization or a regulator can employ to ensure that data processing – whatever the technologies used – is accountable and that when errors occur, as they inevitably will, they are detected and remediated quickly? This question is especially important given the scale at which those processes will need to operate. The goal is not to determine whether one particular application of AI is in compliance and fair at one moment in time, but rather to know that all applications are being examined and monitored on an ongoing basis with the overarching objective of continuous improvement and risk mitigation. Therefore, a key focus of both organizations and regulators should be on developing, assessing, and improving the processes for doing so.

Processes are useful and necessary in the design, development, and deployment stages of AI. For example, Axon's decision[44] to not use facial recognition in police vest cameras was the result of an ethics review process implemented in the research stage of product development. The technology was not deployed due to the discovery of ethical concerns around bias and inaccuracy that could not be satisfactorily mitigated. This is a potent example of the value of Data Review Boards, discussed in greater detail below, to help ensure not merely legal compliance, but that the use of an AI tool is continuously responsible, appropriate, and consistent with an institution's values.

In cases where the technology is deployed, processes will be necessary to remedy wrongful decisions, provide transparency, and ensure fairness. The more critical the impact of the decision, the greater the need for immediate or instantaneous processes to remedy it.

One way for data protection regulators to improve processes may be to engage with leaders and AI engineers in industry and governments to jointly develop outcome-based guidance and use cases that can then be incorporated into organizational processes. Similar to the High Level Expert Group on AI Guidelines[45] or the Singapore Model AI Governance Framework,[46] regulators providing guideposts to organizational processes can foster responsible and accountable AI deployment while still allowing for innovation both in technology and in the processes used to achieve data protection. Another notable example comes from the UK ICO, which has used an innovative and engaging methodology to develop its AI Auditing Framework by publishing a series of blogs, inviting comments from cross-functional groups of experts and engaging AI technology experts in working on solutions.[47]

Nearly every issue focused on throughout this report has both technological and procedural tools to aid in mitigating tensions, and robust processes to implement these tools are critical. Developing the appropriate processes throughout the product lifecycle and in a cross-functional and organization-wide manner helps promote the development of human-centric AI, build trust in AI systems, and generally helps organizations become better data stewards.

The point here is that the most successful approaches to addressing data protection challenges presented by AI to date have focused not on determinations about specific technologies or applications, but rather on on-going processes to identify, prevent, and mitigate harmful impacts. These processes serve as safeguards and are increasingly needed throughout the product or service lifecycle to ensure fairness, transparency, and other goals of data protection.

C. *A Risk-based Approach to AI*

When assessing the data protection challenges presented by AI applications, it is useful, and indeed consistent with the expectations of most individuals, to consider the potential impact and any risk of harms of the proposed processing on individuals, as well as the risk of not using information. This risk-based approach has been suggested by the Singapore Model AI Governance Framework,[48] the GDPR,[49] and most recently, the US Office of Management and Budget's (OMB) Guidance for Regulation of AI Applications.[50] Uses of AI that pose little risk of harm to individuals, either because the decision being made is unimportant or because the likelihood of a harmful outcome is remote, may understandably warrant less scrutiny. The use of AI to recommend songs or movies, for example, most likely warrant less attention than AI applications used in cars to avoid hitting pedestrians or other vehicles.

The focus on impacts and risks to individuals does not diminish the obligation to comply fully with data protection law, but it can help determine the allocation of scarce resources by both organizations and regulators; it can help assure that appropriate attention is paid to those uses of data that pose greater risks; it can help justify the use of more burdensome or time-consuming mitigation processes when the potential harmful outcomes warrant it; and it can help determine the precautionary or remedial processes that should be in place.

While traditional data protection principles serve the goal of limiting data impacts to individuals, new technologies make it increasingly critical to consider each specific use case and evaluate data processing impacts. "The nature of the AI application and the context in which it is used, define to a great extent which tradeoffs must be made in a specific case . . . AI applications in the medical sector will partly lead to different questions and areas of concern than AI applications in logistics."[51] These tradeoffs may vary by sector, but they more accurately vary by the impact to individuals. For example, the prospective impact of using data to train AI is lower than the impact of using data to make a decision, and the processes in place to protect individuals should reflect that difference.

While there may be multiple ways to frame an assessment of risks and impacts, the figure below (Figure 2) captures one way to analyze a proposed data use. The two primary factors are 1) the sensitivity of data and 2) the degree of impact on individuals from use of such data. These factors can help organizations determine the level of process needed based on a particular context. While the assessment depicted by Figure 2 may prove useful to determine the level of process needed, it should go hand in hand with a more holistic risk evaluation given that risk may depend on many other factors irrespective of the nature of data.

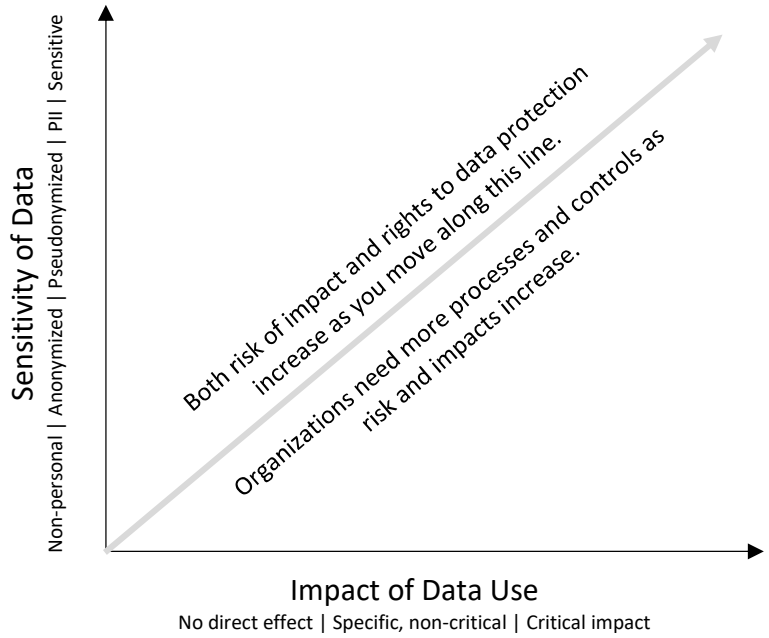

Figure 2. The graph to the left demonstrates one conceptualization of how to analyze risks and impacts of a proposed data use. This is a representation of how to implement a solution that is technology-neutral, impact-focused, and process-oriented.

Relevant questions to consider impacts include: What is the goal of using AI for this application? Is the data used to train an algorithm or deploy it? Is the algorithm making a decision or recommendation? As the sensitivity of data and impact of decision-making increase, organizations should be developing additional processes and controls to limit harmful impacts.

Our discussions have revealed that many organizations are developing tools to understand and mitigate the impacts and harms to individuals from specific AI applications. They range from DPIAs, as required under the GDPR, to specific AI Impact Assessment tools. Some organizations report that they were able to leverage an existing DPIA as a useful tool and process to build upon and perform a wider AI impact assessment.

An important facet of emphasizing data impacts and conducting DPIAs or AI Impact Assessments is to develop a framework to more accurately and consistently assess the impact or harm of a particular data use. There is a need to further develop a better understanding of harms – particularly the potential non-material harms that may occur when collecting and processing data. Analyzing the risk of deploying new models requires understanding how to assess and measure harm and its likelihood of materializing in these new contexts, ranging from monetary harms to non-physical harms such as privacy, security, and discriminatory impacts, among others.[52] CIPL has written in detail about this issue as part of its work on risk mitigation in data protection.[53]

D. *The Need for Data Stewardship and Organizational Accountability*

The rapid and widespread development of new technologies – including AI – has created a renewed need for greater organizational accountability and data stewardship. Developing an impact- and process-oriented approach to data protection will necessarily require that organizations become better data stewards. This will include the need for organizational risk management, improved processes, and better transparency. Data stewardship can be achieved through a number of practices and tools but generally will help instill responsible practices in AI development and deployment.

An important aspect of data stewardship will involve organizations developing principles and values around AI. Responsible data stewardship begins with leadership from the top in an organization. As noted by McKinsey Analytics, "CEOs should make clear exactly what the company goals and values are in various

contexts, ask teams to articulate values in the context of AI, and encourage a collaborative process in choosing metrics."[54] By having leadership articulate the goals and values that an organization is trying to uphold, it allows data stewardship to be part of the corporate culture.

An enhanced focus on data stewardship and organizational accountability is especially necessary in the context of AI. This is because of the challenges in providing individuals with meaningful disclosures about AI tools and algorithms that are difficult even for experts to understand. While a stewardship focus does not eliminate the need for disclosure and transparency, it recognizes that organizations have an obligation to make more thoughtful decisions, and to assume greater responsibility for the consequences of the products, services and technologies that they are developing, in situations where individuals are less able to make informed decisions of their own.

E. *Focusing on Meaningful Roles for Humans*

Some data protection laws appear to contemplate humans as a brake or restraint on automation, at least when automation is used to make decisions that legally or similarly significantly affect individuals. In the case of AI, this runs the risk of never moving beyond the capacity of human minds to be fair, consistent, and rational. We should be more ambitious and aspire for AI to achieve more than human brains can, but this will require considering broader roles for humans within the entire lifecycle of an AI application – from its development to its deployment. Such human involvement must be meaningful and will likely include human oversight over the design, development and deployment process, whether they are building algorithms, evaluating data quality, or testing outcomes, [55] as well as human oversight over the redress process. Above all, we must ensure that this role for human involvement goes beyond compliance checkboxing.

Redress is likely to assume new importance in the effective governance of AI, and therefore should warrant renewed attention. Even with the proper controls and constraints on algorithms, we will never achieve the full potential of AI while also preventing all bad outcomes or even all harms. Rather than viewing the potential risk as a reason for shying away from these new technologies, we should instead strive to ensure that, particularly in the context of automated decision-making with a legal or similarly significant impact, individuals have an effective and efficient avenue for contesting outcomes and appealing decisions.[56] Doing so will help protect not only data protection, but other aspects of human dignity.

F. *Wide Range of Available Tools*

There are a wide range of tools available for organizations looking to improve processes around AI development and deployment. These tools can help organizations facilitate accountability and responsibility in their approach to new technologies and new uses of data. As organizations continue to innovate and improve processes to create new methods for upholding data protection, the tools outlined below reflect some of the current best practices for responsible data users.

1. ***CIPL Accountability Wheel***: The CIPL Accountability Wheel has been used to promote organizational accountability in the broader context of building, implementing and demonstrating comprehensive privacy programs. During our discussions, it became clear that this framework can also be used to help organizations develop, deploy and organize robust and comprehensive data protection measures in the AI context and also to demonstrate accountability in AI. The Accountability Wheel provides a uniform architecture with seven elements for organizations to build and demonstrate their

accountability: Leadership and Oversight; Risk Assessment; Policies and Procedures; Transparency; Training and Awareness; Monitoring and Verification; and Response and Enforcement.[57]

Organizational efforts to promote trustworthiness around AI can map to this wheel to ensure a holistic approach, as each element provides important protections for individuals. The table in Appendix B details examples of some of the existing practices CIPL members are building and implementing to promote organizational accountability as they develop and deploy AI technologies. This is a useful starting point for organizations developing their privacy compliance programs and new policies for AI development, deployment, or use. It is also a useful assessment tool for established organizations to verify that their current practices are comprehensive and efficient.

2. ***AI Data Protection Impact Assessments (AI DPIAs)***: Related to the Risk Assessment element in the above CIPL Accountability Wheel and one of the more common ways to assess the impact of a proposed data use is through a DPIA, which is required under the GDPR for an automated decision that produces legal or similarly significant effects.[58] Many organizations are using DPIAs to comply with data protection and demonstrate their compliance, and some have decided to use DPIAs in an even broader context than that required by law, partially to foster privacy by design and risk mitigation and partially to establish a common lexicon and methodology for assessing data uses across departments and geographies. These assessments may have additional value in the context of AI, and some organizations are developing AI-specific DPIAs, either as a supplement to the assessments required under the GDPR or as an entirely separate assessment.

   AI DPIAs (also referred to as AI Impact Assessments or Algorithmic Impact Assessments) can structure the way that organizations approach assessing issues of fairness, human rights, or other

considerations in these new technologies.[59] The UK ICO's recently issued guidance on DPIAs in the AI context emphasized the fact that a DPIA should be a living document and part of an ongoing organizational process.[60] Singapore's Model Framework, while not calling for DPIAs specifically, also emphasizes the benefit of continual risk assessments to "develop clarity and confidence in using the AI solutions" as well as to help "respond to potential challenges from individuals, other organizations or businesses and regulators."[61]

These assessments can help organizations to build corporate values into their processes, and they will eventually provide a framework of pre-approved cases to set guides for assessing future innovations. Although these sorts of assessments are in their early development, they have the potential to foster fairness, drive accountability, and create consistency. Additionally, AI DPIAs may help organizations develop the data and documentation needed to provide effective transparency to individuals and regulators.

3. ***Data Review Boards (DRBs)***: Data review boards are another potential tool for organizations to structure how they conduct the balancing of interests between the impact of data uses and new AI applications. They also fall under the rubric of "Leadership and Oversight" in the above Accountability Wheel. As with AI DPIAs, DRBs can help organizations respond to new technologies and develop precedent for future ones.[62] A number of organizations already have or are considering DRBs or similar internal or external ethics or AI committees as a way to ensure a human-centric approach to decision-making around AI applications and new data uses. DRBs can help drive organizational accountability, foster responsible decision-making, and ensure that new data uses uphold corporate and societal values.

   Although the structure, procedures, and function of each DRB will be different, there are some best practices that could help organizations ensure the effectiveness of their operation. For example, in terms of the composition of the board, an ideal DRB would have a range of external perspectives – at a minimum from outside the project creators; if possible, from outside the organization entirely, with proper processes in place to account for and protect confidentiality and trade secrets. By having diverse perspectives considering particularly challenging data uses from experts that are to the extent possible detached from commercial interests, DRBs will be positioned to provide meaningful analysis of the issues. Other things to consider include making sure that management is providing adequate support to the DRB, giving the DRB a structural role in the assessment process, and creating tools to ensure that the DRB follows a structured process in an efficient, transparent, and relatively quick manner.

   Creating a DRB will require an organization to consider and document organizational values that the DRB is tasked with upholding. This will foster better decision-making and accountability around many of the important tensions between AI and data protection. For example, DRBs can be useful for assessing and ensuring fairness as well as considering the risks and impacts of new uses or purposes of data. AI DPIAs may be a tool used by the DRB during their assessment of new data uses, or they may be what triggers an issue to be sent to the DRB for further consideration. They can also be consulted as part of the DPIA process itself and provide their views on the risk of a particular processing (see, for instance, Article 35(9) of the GDPR requiring the controller to seek the views of data subjects or their representatives where appropriate – in the same manner, the DRB could provide a different and external perspective).

4. ***Avenues of Redress:*** Many of the concerns around fairness to consumers can be addressed in large part by providing rapid and effective redress through organizational accountability. Redress allows individuals to contest and change an outcome they believe is inaccurate, unfair, or otherwise inappropriate.[63] Depending on the circumstances and impact of the decision, the speed and nature of redress will differ. For example, if a machine is checking boarding passes to allow passengers onto a flight and prevents an individual from boarding, effective redress will need to be nearly instantaneous. More trivial decisions may not need to be explained instantaneously, and redress could be as simple as using an email or logging on to a platform to prompt review of the decision. In either case, the avenue of redress must be visible and accessible to the individuals affected by the decision.

   Redress is often in the form of human review, but there are other forms of redress that can be useful. For example, many smartphones and laptops are using biometric data to recognize approved users. If that technology is not working properly, consumers often have another technological way to bypass it – typically by providing a passcode that the user will have programmed.

   When organizations are developing new technologies and considering the impact of those technologies, it is unlikely that they will foresee and limit every negative impact. In some cases, an organization may determine that the risks are too high to deploy the technology. However, the tradeoffs in other contexts may warrant that the technology is deployed, but that it has visible and effective avenues to correct situations where biased or incorrect decision-making occurs. Organizations should ensure that redress is meaningful – and that it does not merely become a rubber stamp on an automated decision. If unfairness or inaccuracy is uncovered, organizations should have processes in place for limiting similar situations in the future. Considering and developing these remedies and processes will be an essential part of deploying AI, and regulators evaluating the use of AI and impact on data protection should look for these visible avenues of redress as one way to demonstrate responsible implementation of AI technologies.

IV. Conclusion

The technologies of AI, the volume and variety of AI tools, and the speed with which they are evolving and being deployed present many challenges to data protection. AI can include automated decisions, but it can also include augmenting human intelligence to produce better outcomes. The numerous benefits produced by the proliferation of AI technologies are not without challenges. However, a year's worth of roundtables, discussions, and research have clearly shown that there is both sufficient flexibility in most data protection laws and sufficient creativity among organizations and regulators to comply with those laws and to ensure that AI is developed and deployed in ways that are not merely lawful, but beneficial and accountable.

Appendix A. *Translations of Fairness* – The following table provides the translation of fairness, as stated in Article 5, in each of the 23 languages with an official translation of the GDPR.

Language	**Article 5 wording**	**Google Translation**
Bulgarian	добросъвестност	Good faith
Croatian	poštenosti	Honesty
Czech	korektnost	Correctness, propriety
Danish	rimelighed	Reasonably
Dutch	behoorlijkheid	Goodness
English	fairness	Fairness
Estonian	õiglus	Justice, justness, fairness, equity, righteousness
Finish	kohtuullisuus	Equity, fairness
French	loyauté	Loyalty, trustworthiness, fidelity
Gaelic	cothroime	Fairness
German	Verarbeitung nach Treu und Glauben	Good faith processing
Greek	αντικειμενικότητα	Objectivity
Hungarian	tisztességes eljárás	Due process, fair play
Italian	correttezza	Correctness, fairness, propriety, honesty
Latvian	godprātība	Integrity, honesty, good faith
Lithuanian	sąžiningumo	Fairness, honesty, integrity, good faith
Maltese	ġustizzja	Justice, fairness
Polish	rzetelność	Reliability, dependability, honesty, rectitude, squareness
Romanian	echitate	Fairness, equity, justice, uprightness
Slovak	spravodlivosť	Justice, justness, equity, rectitude, uprightness, narrow way, virtuousness
Slovenian	pravičnost	Justice
Spanish	lealtad	Loyalty, allegiance, devotion
Swedish	korrekthet	Correctness, propriety

Appendix B. *Mapping Best Practices in AI Governance to the CIPL Accountability Wheel*

Throughout the roundtables, organizations offered some of the practices they use to ensure responsible and accountable deployment of AI. The table below surveys some of those practices.

Accountability Element	Related Practices
Leadership and Oversight	• Public commitment and tone from the top to respect ethic, values and specific principles in AI development • Institutionalized AI processes and decision-making • Internal Code of Ethics rules • AI/ethics/oversight boards, councils, committees (internal and external) • Appointing board member for AI oversight • Appointing responsible AI lead/officer • Privacy/AI engineers and champions • Set up an internal interdisciplinary board (lawyers, technical teams, research, business units) • Appointment of privacy stewards to coordinate others
Risk Assessment	• Understand the AI purpose and use case in business and processes – for decision-making, or to input into decision, or other • Understand impact on individuals • Fairness assessment tools • Algorithmic Impact Assessment • Ethics Impact Assessment • Broader Human Rights Impact Assessment • DPIA for high risk processing • Consider anonymization techniques • Document tradeoffs (e.g. accuracy vs. data minimization, security vs. transparency, impact on few vs. benefit to society)
Policies and Procedures	• High level principles for AI – how to design, use, sell, etc. • Assessment questions and procedures • Accountability measures for two stages – training and decision-making • White, black, and gray lists of AI use • Evaluate the data against the purpose – quality, provenance, personal or not, synthetic, in-house or external sources • Verification of data input and output • Algorithmic bias – tools to identify, monitor and test and including sensitive data in datasets to avoid bias • Pilot testing AI models before release • Use of encrypted data or synthetic data in some AI/ML models • Use of high quality but smaller data sets • Federated AI learning models (data doesn't leave device) • Special considerations for companies creating and selling AI models, software, applications • Due diligence checklists for business partners using AI tech and tools • Using external tools, guidelines, self-assessment checklists

Transparency	• Different needs for transparency to individuals, regulators, business and data partners and internally to engineers and leadership • Explainability is part of transparency and fairness • Transparency trail – explainability of decision and broad workings of algorithm + more about the process than the technology + what factors + what testing to be fair + accountability for impact of decisions on a person's life + what extent of human oversight • Explain that it is an AI/ML decision, if there is the possibility of confusion (Turing test) • Provide counterfactual information • Differentiated and flexible transparency – linked to context, audience/users, purpose of explainability and risk, severity of harm – prescriptive lists of transparency elements is not helpful • Factsheets and Model Cards • Understand customers' expectations and deploy based on their readiness to embrace AI – tiered transparency • From black box to glass box – looking at the data as well as the algorithm /model; aspiration of explainability helps understand the black box and builds trust
Training and Awareness	• Data scientist training, including how to avoid and address bias • Cross functional training – privacy professionals and engineers • Ad hoc and functional training • Fairness training • Ethics training • Use cases where problematic AI deployment has been halted • Role of "Translators" in organizations, explaining impact and workings of AI
Monitoring and Verification	• Purpose of AI determines how much human intervention is required • Human in the loop – in design, in oversight, in redress • Human understanding of the business and processes using AI • Human development of software and processes • Human audit of input and output • Human review of individual decisions • Ongoing monitoring, validation and checks • Oversight committees even in design stage • Redress to a human, not to a bot • Monitoring the eco-system from data flow in, data process and data out • Reliance on different audit techniques • Version control and model drift, tracking of black box, algorithms by engineers • RACI models for human and AI interaction
Response and Enforcement	• Complaints-handling • Redress mechanisms for individuals to remedy AI decision • Feedback channel • Internal supervision of AI deployment

References

[1] "First Report: Artificial Intelligence and Data Protection in Tension," CIPL (10 October 2018), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf, at page 19.

[2] The events included CIPL/Singapore Personal Data Protection Commission (PDPC) Joint Interactive Working Session on "Accountable and Responsible AI" (16 November 2018, Singapore); Roundtable with European Regulators and Industry Participants on "Hard Issues in AI" (12 March 2019, London); Roundtable with the EU Commission High Level Expert Group on "Ethics Guidelines for Trustworthy AI" (27 June 2019, Brussels); Roundtable with Asian Regulators and Industry Participants on "Personal Data Protection Challenges and Solutions in AI" (18 July 2019, Singapore); Roundtable with the UK Information Commissioner's Office (ICO) on "AI Auditing Framework" (12 September 2019, London); CIPL/TTC Labs Design Jam on "AI Explainability" (3 December 2019, Cebu); and CIPL Industry Workshop on a European AI Regulatory Approach (14 January 2020, Brussels).

[3] For example, GDPR, Article 5 "Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');" New Zealand Privacy Act, Section 6, Principle 4(b)(i) "Personal information shall not be collected by an agency (b) by means that, in the circumstances of the case (i) are unfair;" Draft India Personal Data Protection Law, Section 5(a) "Every person processing personal data shall process such personal data (a) in a fair and reasonable manner and ensure the privacy of the data principle."

[4] See "FTC Policy Statement on Unfairness" (17 December 1980), available at https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness.

[5] European Data Protection Board, Guidelines 4/2019 on Article 25: Data Protection by Design and by Default (adopted on 13 November 2019) available at https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf.

[6] There have been other attempts to describe or define fairness. The UK ICO, for example, described fairness as meaning that "you should only handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them." "Principle (a): Lawfulness, Fairness, and Transparency," UK ICO, available at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/. The GDPR also gives some guidance regarding fairness or how to achieve fairness, for example, in Recital 71, which states "In order to ensure fair and transparent processing in respect of the data subject… the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect."

[7] For background information, see Douglas MacMillan and Nick Anderson, "Student Tracking, Secret Scores: How College Admissions Offices Rank Prospects Before They Apply," Washington Post (14 October 2019), available at https://www.washingtonpost.com/business/2019/10/14/colleges-quietly-rank-prospective-students-based-their-personal-data/?fbclid=IwAR24p1HKEaHfN0K7kH4H5XBeDw4qgRib_v-o48afJ5bF5z10dIe9vCtiVac.

[8] "In a move rarely seen among tech corporations, [Axon] convened the independent board last year to assess the possible consequences and ethical costs of artificial intelligence and facial-recognition software. The board's first report, published Thursday, concluded that 'face recognition technology is not currently reliable enough to ethically justify its use' — guidance that Axon plans to follow." Deanna Paul, "A Maker of Police Body Cameras Won't Use Facial Recognition Yet, for Two Reasons: Bias and Inaccuracy" (28 June 2019), available at https://www.washingtonpost.com/nation/2019/06/29/police-body-cam-maker-wont-use-facial-recognition-yet-two-reasons-bias-inaccuracy/.

[9] Elizabeth Denham, "Blog: Live Facial Recognition Technology – Police Forces Need to Slow Down and Justify Its Use," UK ICO, available at https://ico.org.uk/about-the-ico/news-and-events/blog-live-facial-recognition-technology-police-forces-need-to-slow-down-and-justify-its-use/.

[10] "Ethics Guidelines for Trustworthy AI," European High-Level Expert Group (HLEG) on Artificial Intelligence (8 April 2019), available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, at page 30.
[11] "Treating Customers Fairly – Towards Fair Outcomes for Consumers," United Kingdom Financial Services Authority (July 2006), available at https://www.fca.org.uk/publication/archive/fsa-tcf-towards.pdf, at pages 11-13.
[12] Singapore Personal Data Protection Commission, "A Proposed Model Artificial Intelligence Governance Framework," (January 2019), available at https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf, at page 15. "A decision is fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group."
[13] *Supra* note 1, at page 18.
[14] Certain data protection concepts are in tension with the techniques used to identify the risk of bias. For example, although GDPR Article 9 prohibits processing of certain special categories of personal data unless certain exceptions apply, collecting racial or ethnic origin data is likely necessary to identify whether an algorithm is discriminating by ethnicity or race. Although data protection principles hope to protect individuals by limiting collection of such information, this data is critical to achieve the goal of mitigating and detecting bias.
[15] Google, Perspectives on Issues in AI Governance (2019), available at https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf, at page 15.
[16] Rumman Chowdhury, "Tackling the Challenge of Ethics in AI," Accenture (6 June 2018), available at https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai.
[17] Kush R. Varshney, "Introducing AI Fairness 360," IBM Research Blog (19 September 2018), available at https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/.
[18] "Manage AI, with Trust and Confidence in Business Outcomes," IBM, available at https://www.ibm.com/downloads/cas/RYXBG8OZ.
[19] GDPR, Article 35(1).
[20] "International Resolution on Privacy as a Fundamental Human Right and Precondition for Exercising Other Fundamental Rights," 41st International Conference of Data Protection & Privacy Commissioners (ICDPPC) (21-24 October 2019), available at https://privacyconference2019.info/wp-content/uploads/2019/10/Resolution-on-privacy-as-a-fundamental-human-right-2019-FINAL-EN.pdf.
[21] *Supra* note 15, at page 13.
[22] *Supra* note 1, at page 15; See also GDPR, Article 12.
[23] See, for example, GDPR, Articles 5 & 12; Brazil LGPD, Articles 14(2) and (6), Article 18 (2), (7) and (8); and Draft India Personal Data Protection Law, Section 23.
[24] "Project ExplAIn: Interim Report," UK Information Commissioner's Office (3 June 2019), available at https://ico.org.uk/media/2615039/project-explain-20190603.pdf, at page 15.
[25] GDPR, Article 22.
[26] *Supra* note 10, at page 20.
[27] *Id*.
[28] *Supra* note 12, at page 13. The HLEG Guidelines for Trustworthy AI categorizes transparency considerations similarly to the Singapore Model Framework, focusing on traceability, explainability, and communication.
[29] "Like nutrition labels for foods or information sheets for appliances, factsheets for AI services would provide information about the product's important characteristics." Aleksandra Mojsilovic, "Factsheets for AI Services," IBM Research Blog (22 August 2018), available at https://www.ibm.com/blogs/research/2018/08/factsheets-ai/.
[30] One best practice could be the use of Model Cards, which are "short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions . . . that are relevant to the intended application domains." Margaret Mitchell et al., "Model Cards for Model Reporting" (14 January 2019), available at https://arxiv.org/pdf/1810.03993.pdf. See also "The Value of a Shared Understanding of AI Models," Google, available at https://modelcards.withgoogle.com/about.
[31] GDPR, Article 5(1)(b) [emphasis added].
[32] "The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose." OECD Revised Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2013), available at http://oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.

[33] See for example, GDPR, Article 6(4). Note, however, that the GDPR prohibits using personal data for a purpose other than for which it was originally collected, unless the new purpose is "not incompatible" with the original.
[34] GDPR, Article 6(4). The criteria for further compatible processing under the GDPR include (a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing; (b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller; (c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10; (d) the possible consequences of the intended further processing for data subjects; and (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.
[35] GDPR, Article 6(4)(d).
[36] In the EU, this could potentially be considered under the umbrella of "research purposes." GDPR, Article 5(1)(b) (allowing further processing for "archiving purposes in the public interest, scientific or historical research purposes or statistical purposes").
[37] Fred H. Cate and Rachel D. Dockery, "Artificial Intelligence and Data Protection: Observations on a Growing Conflict," *Seoul National University Journal of Law & Economic Regulation*, Vol. 11. No. 2 (2018), at page 123.
[38] GDPR, Article 5(1)(c).
[39] Potential alternatives to using personal data include using synthetic data sets, differential privacy, or other privacy enhancing techniques. For considerations related to data minimization and AI, see the UK Information Commissioner's Office blog on its AI Auditing Framework, Rueben Binns and Valeria Gallo, "Data Minimisation and Privacy-Preserving Techniques in AI Systems," UK ICO (21 August 2019), available at https://ai-auditingframework.blogspot.com/2019/08/data-minimisation-and-privacy_21.html.
[40] See, e.g., *supra* note 1 ("AI, and the variety of data sets on which it often depends, only exacerbates the challenge of determining when data protection laws apply by expanding the capability for linking data or recognising patterns of data that may render non-personal data identifiable…. Simply stated, the more data available, the harder it is to de-identify it effectively.").
[41] See, e.g., Ursula von der Leyen, "A Union that Strives for More: My Agenda for Europe, Political Guidelines for the Next European Commission," available at https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf, at page 13 ("In my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence."). Although, this initiative may ultimately take the form of a policy paper that will spell out different options for a legal framework on AI which may lead to formal proposals later in the year.
[42] "AI technology needs to continue to develop and mature before rules can be crafted to govern it. A consensus then needs to be reached about societal principles and values to govern AI development and use, followed by best practices to live up to them. Then we're likely to be in a better position for governments to create legal and regulatory rules for everyone to follow." The Future Computed, Microsoft (2018), available at https://blogs.microsoft.com/wp-content/uploads/2018/02/The-Future-Computed_2.8.18.pdf, at page 9.
[43] "Artificial Intelligence Impact Assessment," Platform for the Information Society (2018), available at https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf, at page 21.
[44] *Supra* note 8.
[45] *Supra* note 10.
[46] *Supra* note 12.
[47] AI Auditing Framework, UK ICO, available at https://ico.org.uk/about-the-ico/news-and-events/ai-auditing-framework/.
[48] *Supra* note 12, at page 6-7.
[49] This is apparent in the DPIA requirement in Article 35, among others. For an overview of risk-based provisions of the GDPR, see Gabriel Maldoff, "The Risk-Based Approach in the GDPR: Interpretation and Implications," International Association of Privacy Professionals, available at https://iapp.org/media/pdf/resource_center/GDPR_Study_Maldoff.pdf.
[50] "[A] risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits. Agencies should be transparent about their evaluations of risk and re-evaluate their assumptions and conclusions at appropriate intervals so as to foster accountability." Russel Vought, "Draft Memorandum for the Heads of Executive

Departments and Agencies: Guidance for the Regulation of Artificial Intelligence Applications," US Office of Management and Budget (7 January 2019), available at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

[51] *Supra* note 43, at page 8.

[52] For one example of an effort to assess such harms, it may be useful to look at the UK's Online Harms White Paper, which is part of an ongoing consultation to analyze online harms and implement safety measures. "Online Harms White Paper," Department for Digital, Culture, Media & Sport (April 2019), available at https://www.gov.uk/government/consultations/online-harms-white-paper.

[53] "Paper 1: A Risk-based Approach to Privacy: Improving Effectiveness in Practice," CIPL, (19 June 2014), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/white_paper_1-a_risk_based_approach_to_privacy_improving_effectiveness_in_practice.pdf; "Paper 2: The Role of Risk Management in Data Protection," CIPL (23 November 2014), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/white_paper_2-the_role_of_risk_management_in_data_protection-c.pdf; "Protecting Privacy In a World of Big Data: Paper 2: The Role of Risk Management," CIPL (16 February 2016), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/protecting_privacy_in_a_world_of_big_data_paper_2_the_role_of_risk_management_16_february_2016.pdf; "Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR," CIPL (21 December 2016), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf.

[54] Roger Burkhardt, Nicolas Hohn, and Chris Wigley, "Leading Your Organization to Responsible AI," McKinsey Analytics (May 2019), available at https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Leading%20your%20organization%20to%20responsible%20AI/Leading-your-organization-to-responsible-AI.ashx, at pages 3-4.

[55] Some have described the level of human involvement with the terms human-in-the-loop, human-out-of-the-loop, and human-over-the-loop. Human-in-the-loop refers to situations where "human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input;" human-out-of-the-loop refers to situations where "there is no human oversight over the execution of decisions" and "AI has full control without the option of human override;" and human-over-the-loop "allows humans to adjust parameters during the execution of the algorithm." See *supra* note 12, at page 8.

[56] This sort of human review already exists in US law in certain contexts where automated processing informs decision-making. The Fair Credit Reporting Act and Equal Credit Opportunity Act both provide consumers with some right to explanation and to contest the decision, while Title VII of the Civil Rights Act as well as the Fair Housing Act provide individuals with a right to challenge decisions.

[57] "The Case for Accountability: How It Enables Effective Data Protection and Trust in the Digital Society," CIPL (23 July 2018), available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_1_-_the_case_for_accountability_-_how_it_enables_effective_data_protection_and_trust_in_the_digital_society.pdf.

[58] GDPR, Article 22; see also GDPR, Article 35(1) ("Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data.").

[59] One potential roadmap for AI DPIAs is to "1. Map the (public) benefits of an AI application. 2. Analyse the reliability, safety and transparency of AI applications. 3. Identify values and interests that are concerned by the deployment of AI. 4. Identify and limit risks of the deployment of AI. 5. Account for the choices that have been made in the weighting of values and interests." See *supra* note 43, at page 25.

[60] The ICO guidance also suggested 5 elements for AI DPIAs, including 1) a systemic description of the processing, 2) assessing necessity and proportionality, 3) identifying risks to rights and freedoms, 4) measures to address the risks, and 5) a 'living' document. Simon Reader, "Data Protection Impact Assessments and AI," UK ICO AI Auditing Framework (23 October 2019), available at https://ai-auditingframework.blogspot.com/2019/10/data-protection-impact-assessments-and.html.

[61] *Supra* note 12, at page 8 (though these are called "risk assessments" in this framework).

[62] Accenture has recently published a white paper detailing the benefits of data and AI ethics committees, exploring different setups and best practices for them. The paper explains that "to be successful, it must be thoughtfully designed, adequately resourced, clearly charged, sufficiently empowered, and appropriately situated within the organization." John Basl and Ronald Sandler, "Building Data and AI Ethics Committees," Accenture & Northeastern University Ethics Institute (2019), available at https://www.accenture.com/_acnmedia/pdf-107/accenture-ai-data-ethics-committee-report-executive-summary.pdf, at page 2.
[63] *Supra* note 10, at page 20 (When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. Knowing that redress is possible when things go wrong is key to ensure trust.); see also *supra* note 12, at page 17 (discussing the need for feedback channels so that individuals can review and correct their data, as well as decision review channels to contest adverse decisions).