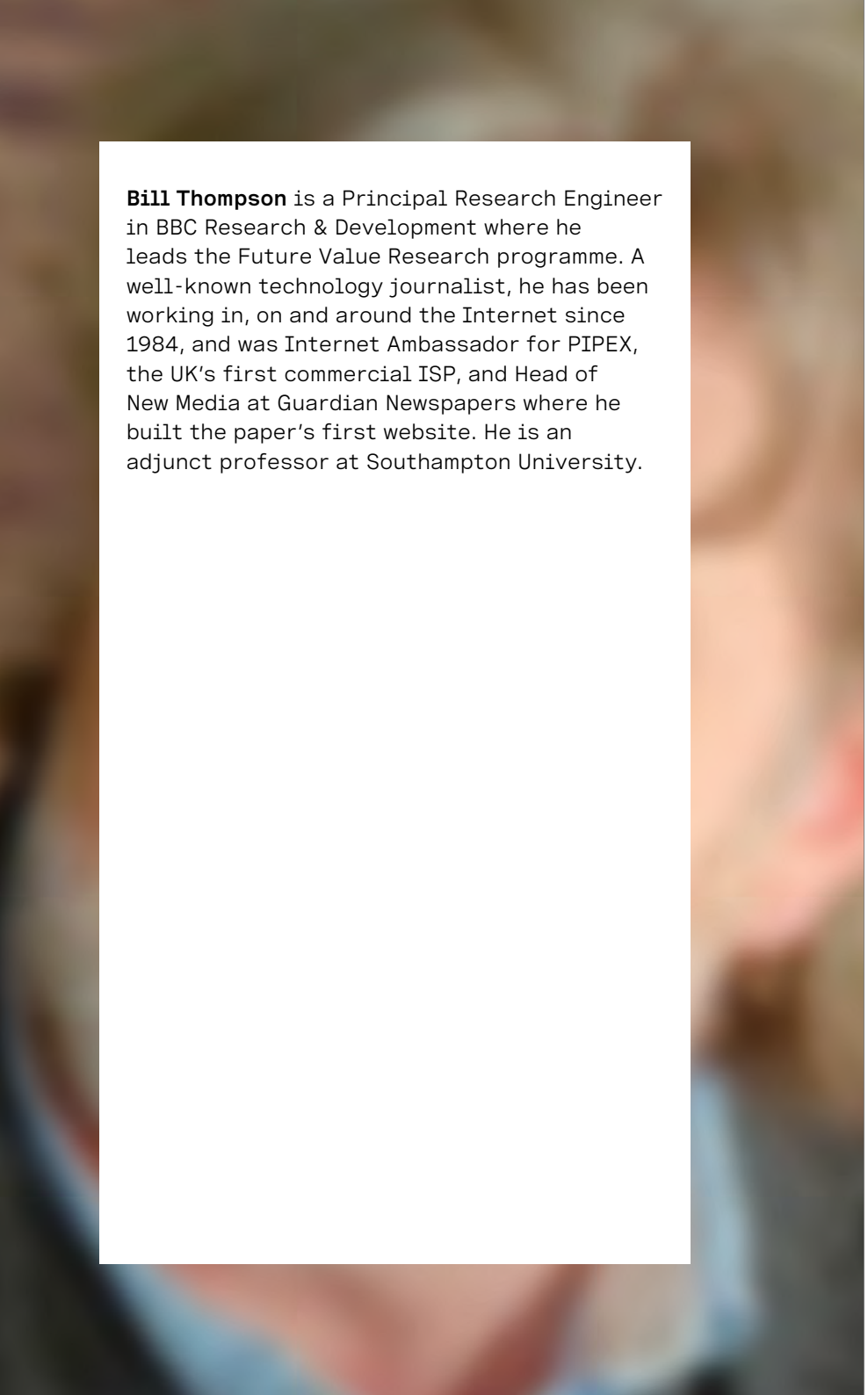

Bill Thompson is a Principal Research Engineer in BBC Research & Development where he leads the Future Value Research programme. A well-known technology journalist, he has been working in, on and around the Internet since 1984, and was Internet Ambassador for PIPEX, the UK's first commercial ISP, and Head of New Media at Guardian Newspapers where he built the paper's first website. He is an adjunct professor at Southampton University.

Data from cradle to grave: How personal data stores could transform the uses of data about children and young people

Bill Thompson, BBC Research & Development

An age of watchers

The growth of 'surveillance capitalism', a term popularised by Shoshana Zuboff in her book *The age of surveillance capitalism* (2019) and now widely adopted, should not surprise us. Computer systems, from the earliest mainframes to the modern pocket-sized networked supercomputers we still call 'phones', have always maintained records of how they are being used, in order to manage performance, monitor security and, where relevant, charge for resources.

Keeping records - or 'logging' - is what computers do, and from the very earliest days of online publishing on the website owners were promised that the ability to track visitors was one of the great advantages. But we have come a long way from using web server logs to let us know how many people read a web page and for roughly how long, and now vast amounts of data are being collected about every computer-related transaction.

The extent of logging and the ways the records are being used is the result of a set of choices and imperatives that have largely been driven by the increased commercialisation of the online environment since the mid-1990s (Naughton, 2012). This complex, expensive, flawed and intrusive system provides more

and more data points around any interaction a user may have with a networked system, processing that data in order to make inferences about them that can be used to drive advertisements, shape online experiences, or even alert the state to activity that it deems noteworthy or dangerous (Zuboff, 2019).

And we are not at the end of this process. As technology develops, so does tracking. With the emergence of virtual reality (VR) and augmented reality (AR) systems and the push to develop shared, persistent, online spaces that allow for property ownership, as multiplayer games fuse into what has been called the 'metaverse', this will only get more extreme. Unless we do something to avoid it, we can look forward to an age of 'omniscient capitalism' (Pesce, 2019), in which everyone who straps on a headset or puts on a pair of smart glasses becomes part of a virtual panopticon, with consequent risks to privacy.

We need to ask ourselves what can be done about this, particularly with regard to data about children and young people.

Data about children and young people

In the ongoing debate about the balance of interests between the technology companies that want to monitor users to support their business models, governments that want to track their citizens, and individuals who want ways to preserve their privacy and exert control over their data (European Parliament, 2022), it is generally recognised that there are special considerations concerning children and young people.

Some of these are legal, relating to their ability to give informed consent, while there are also issues around how much we want the online activities of children and young people to be monitored, profiled and used to present ads, shape what they see on social media, or even influence their life trajectory.

In some cases, they create the problem for themselves. Many young people have found a way to work around the age limits of services like TikTok, Snapchat and Facebook by lying on registration forms, or had accounts created by compliant parents or caregivers so that they are not 'missing out'. They may even be using mobile phones registered to adults, which means that personal data is collected from them as if they are over 18, clicking through consent screens without having the legal authority to accept the terms (Ofcom, 2022).

Even when someone's age is apparent there can be problems, and discussions about an appropriate regulatory environment have started to consider areas where data about children may be slipping through the gaps. In the UK the recent Digital Futures Commission report on data-driven education systems, *Governance of data for children's learning in UK state schools* (Day, 2021), prompted in part by the shift from in-person to online teaching during the COVID-19 pandemic, pointed out just how regulatory uncertainties and common business practices around excessive data protection and retention had created an environment where children's education data was largely uncontrolled, compromising the many potential benefits that data processing could offer.

This issue is not limited to education data. Recently in the USA the Federal Trade Commission (FTC) ordered WW (formerly 'Weight Watchers') to delete a dataset and the training model derived from it because it had been illegally acquired from young people (Brody, 2022). This should not surprise us, as many aspects of young people's lives are now mediated by electronic systems that play such an important part in our lives, and so they have become full members of Zuboff's 'surveillance society'.

Looking beyond regulation

Whatever regulations are in place, data management is fundamentally a technical issue, involving the collection, storage and processing of computerised records. At the moment the standard model for organisations that want to use data about people is for them to set up a central database or user activity store, often linked to user accounts with some form of validation or login capability. Keeping this data secure is a challenge, and there are significant reputational, regulatory and financial risks, as well as the costs of storage and service provision. It is also very easy for data to be used in ways that go beyond the original purposes, with potential privacy implications.

Away from this standard approach there is much

experimentation and innovation around data management. One of the most promising alternatives to the monolithic model is the personal data store, or PDS. A PDS is a storage system in which data can be securely stored, using a range of encryption and other technologies. Some are 'data vaults', like Solid[1] or Mydex[2]. Others like Databox[3] are complete computing environments where data is both stored and processed without leaving the secure area, with only the results made available to third parties. The owner of a PDS can store their personal data and control which systems have access to it, changing their mind at any time.

The PDS is an enabling technology that is capable of supporting a wide range of business models, from open source to fully commercial, with a range of tools to facilitate the integration of third-party apps (Bolychevsky & Worthington, 2018), but the technology is mature enough for the Flanders government to have started a project to give every citizen a Solid PDS to hold citizen data (Berners-Lee & Bruce, 2021).

The PDS is an element in a broader effort to rethink data processing around the trusted processing of data, and a number of organisations, including the BBC, have developed a model they call the 'public service data ecosystem' (PSDE), 'a set of components which work together to provide a secure and effective platform for public service applications, and which are able to integrate personal data with open data, aggregate data, and data from sources such as Internet of Things devices' (Sharp et al., 2021).

The idea of a data ecosystem came from an ongoing discussion over the future of data governance more generally, and detailed exploration of new legal models for data stewardship, where the control of the data about people used by an organisation is managed by a data trust or cooperative that balances the interests of the individuals and the organisation (Ada Lovelace Institute, 2021).

At the heart of the public service data ecosystem is a personal data store that provides a user-centred approach to the storage and access of both legally defined 'personal data' and other data the user might want to control independent of any particular application. Within the public service data ecosystem each user can have one or more data stores, which

can be located in the cloud or on their own hardware, and third parties cannot copy or perform any processing of the data without the user's explicit permission.

The ecosystem also incorporates other data sources, whether open data or licensed, that can be used in combination with the data stored in the PDS to support a range of services. These could include finance, health and entertainment applications, all of which benefit from the additional control over data use that the PDS provides.

PDS-based systems are already widely used. The CitizenMe app is used to support a model called 'zero-party data', holding user data within their app, while allowing it to be used in controlled and transparent ways (Deakins, 2022). Mydex is working with the Scottish Government to provide identity verification services by storing authorised credentials (Mydex, 2020).

In 2021 BBC Research & Development (R&D) published its work with the Solid system to develop a cross-media recommender that used combined transaction data from multiple media services including the BBC iPlayer to create a user profile that could improve the quality of recommendations from those services, without the need to share the data across services (Sharp, 2021; Sharp et al., 2021).

Delivering the potential of the PDS

While we have been talking about PDS in some form for well over a decade, the market has so far failed to deliver on their potential. A study on PDS conducted at the Cambridge Judge Business School (University of Cambridge) in 2015 noted that:

> As an innovative concept, the personal data store faces significant obstacles to widespread diffusion. In particular, PDS providers must reach critical mass in the context of a double-sided market: the PDS system must attract a sufficient number of individuals and businesses if it is to flourish as a platform for data exchange, but neither individuals nor businesses are easily captured without the other first in place. (Brochot et al., 2015)

Things now seem to be changing, partly as a result of legal rulings including the FTC ruling against WW referred to earlier, and notably the Belgian Data Protection Authority's recent decision to fine the online advertiser IAB Europe over its transparency and consent framework (Bryant, 2022), as well as new regulations like the UK Age Appropriate Design Code (AADC). All of these have raised awareness of how data can be abused, and the risks organisations take when they store data.

At the same time, research into attitudes to personal data use has shown that people dislike the current approach in which commercial organisations control their personal data, preferring approaches that give them control over their data that include oversight from regulatory bodies or that enable them to opt out of data gathering (Hartman et al., 2020).

And we now have a population, including young people, which is more familiar with security practices for online services and smartphones, while advertising companies like Apple have raised awareness both of surveillance and how to counter it. Apple advertising makes much of the fact that their photos app processes your images on your phone while Google sends your data to its cloud.

Young people and personal data

The combination of a shifting regulatory environment, increased consumer awareness and technical maturity creates an opportunity to propose PDS-based approaches to the management of the data needed to deliver services to children and young people, offering a large enough market to make it worth investing in, and a compelling use case that can drive providers towards the technology.

A PDS-based approach would let service providers, especially in the educational technology (EdTech) market, offer advanced functionality to schools while protecting user data, and young people could then be encouraged to use their PDS for other purposes, perhaps encouraging social media platforms to offer children-oriented services using the same technology. It does not matter who provides the PDS as long as a service can make use of it through a standard interface with a suitable data model.

Within the broader context of the public service data ecosystem, other data sources, including open data, could be safely combined with data held in a PDS to support a range of services - for example advanced profiling and recommendations, as discussed by BBC R&D (Sharp et al., 2021). Furthermore, the data would remain under the control of the individual and would not have to be deleted when they left school or added to the data lakes maintained by the companies providing services. And, of course, a PDS-based model for EdTech would also be useful for adult learners, so it could be that many would choose to retain their PDS as they leave formal education, putting pressure on platforms and others to adapt to the new model.

Remaining issues

A public service data ecosystem based around PDS does not solve all of the issues around the use and abuse of data about people. There are still consent issues to be dealt with when it comes to children and young people, both over the initial provision of a PDS to someone under 18 and to their agreement to allow data to be accessed by services.

Nor can it deal with the 'selling a kidney' problem, where an individual who has control of all their personal data decides to give access - or a copy - to an organisation without really considering the implications, or on the basis of misleading offers.

There are also, inevitably, going to be security breaches, bugs in code and other ways in which there could be data breaches. As always, we will need a strong regulatory framework, and proper enforcement of penalties against those who abuse data about people, to reinforce the technical provisions. However, an approach to the public service data ecosystem based on personal data stores seems to be healthier and more likely to encourage people to understand and consider the consequences of their actions.

The benefits of control

Perhaps the biggest benefit will be increased transparency, because every application and service that wants to create and use data about a user via a PDS will have to be very clear about what it is storing, and when it wants to use that data.

There are parallels between the current debate concerning the ways we manage data about people, and the ongoing arguments about accessibility of websites and games. Early websites and games were generally designed for users who required no special accommodations (other than vision correction, because somehow, not being able to focus is not classed as a disability). Making the web more accessible and adding accessibility features to games was presented as limiting the creative expression of designers, and was resisted for many years.

There is a similar debate about data processing, where companies argue that regulation or technical limitations inhibit creativity. Yet, as with accessible websites and games, alternative approaches that put respect for people at the centre of their design can still deliver business objectives, perhaps more easily and certainly with less regulatory peril.

Computerised systems will remain important in all aspects of our lives, including in educational settings, and storing and processing data about users is a vital aspect of their operation that cannot be avoided. However, it is possible to develop approaches to data management that allow data about people to be controlled by them while still being available where they are needed to serve legitimate interests. The PDS model offers individuals a significant degree of control over their data without compromising the functionality of the systems, and merits serious consideration as an alternative to large-scale databases.

A world within which children and young people are encouraged to think about their personal data and where and how it is stored and used, with PDS as one option, may be one in which they grow up into adult data citizens instead of being seen as data 'subjects'. In order for this to happen we need further research to explore the capabilities of the technology, and a regulatory environment that can accommodate it as one option for the storage and processing of personal data.

Ada Lovelace Institute. (2021). Exploring legal mechanisms for data stewardship

Berners-Lee, T., & Bruce, J. (2021). Committing to a digital innovation economy in Flanders, built on Solid

Bolychevsky, I., & Worthington, S. (2018). Are personal data stores about to become the NEXT BIG THING? *Medium*, 4 October

Brochot, G., Brunini, J., Eisma, F., Larsen, R., & Lewis, D. (2015). Study on personal data stores. *Digital Single Market*, 7 August

Brody, B. (2022). Weight Watchers must delete algorithms built from kids' data. *Protocol*, 4 March

Bryant, J. (2022). Belgian DPA fines IAB Europe 250K euros over consent framework GDPR violations. Iapp.org, 2 February

Databox Project. (2017). EPSRC Databox Project

Day, E. (2021). *Governance of data for children's learning in UK state schools*. Digital Futures Commission & 5Rights Foundation

Deakins, S. (2022). The future of data is 'zero data'. *CitizenMe*, 15 June

European Parliament. (2022). Deal on Digital Markets Act: Ensuring fair competition and more choice for users. News

Hartman, T., Kennedy, H., Steedman, R., & Jones, R. (2020). Public perceptions of good data management: Findings from a UK-based survey. *Big Data & Society, 7*(1), 205395172093561

Mydex. (2020). *Smart entitlements: Recommendations and report for the Scottish government*

Naughton, J. (2012). *From Gutenberg to Zuckerberg: What you really need to know about the internet*. Quercus

Ofcom. (2022). Children and parents: Media use and attitudes report 2022

Pesce, M. (2019). *Augmented reality: Unboxing tech's next big thing*. Polity

Sharp, E. (2021). Personal data stores: Building and trialling trusted data services. BBC R&D

Sharp, E., Ricklefs, H., Leonard, M., Jones, R., Greenham, A., Carter, J., Broom, T., Bird, K., & Thompson, B. (2021). Enhancing media through the development of a Public Service Data ecosystem. *IBC 2021*

Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.

1 For more information about this product, see: https://solidproject.org

2 For more information about this product, see: https://mydex.org

3 For more information about the architecture of Databox, see https://imperial.ac.uk/systems-algorithms-design-lab/research/databox-project