CENTER FOR
THE GOVERNANCE
OF CHANCE

DATA, PRIVACY, AND THE INDIVIDUAL

THE ETHICS OF DATA ACQUISITION

N. WILDMAN, A. ARCHER, H.M. BROUWER, A.M. CAWSTON
TILBURG UNIVERSITY

NOVEMBER 2019

THE ETHICS OF DATA ACQUISITION: PROTECTING PRIVACY AND AUTONOMY WHILE HARNESSING THE POTENTIAL OF BIG DATA

Dr. N. Wildman, Dr. A. Archer, H.M. Brouwer, Dr. A.M. Cawston

Tilburg Centre for Logic, Ethics, and Philosophy of Science (TiLPS)

Tilburg University

Reference to this paper should be made as follows:
Wildman, N., Archer, A., Brouwer, H.M., Cawston, A.M. (2019) "The Ethics of Data Acquisition: Protecting Privacy and Autonomy While Harnessing the Potential of Big Data", *Data, Privacy and the Individual*.

INTRODUCTION

The sharing of data by individuals and organizations for research, policy-making, and humanitarian purposes is increasingly recognized as a crucial mechanism for building a better society (Chretien et al 2016). However, issues such as a lack of clarity on data control and ownership; disagreement about how best to respect individual rights, consent, and privacy; limited technical understanding; and the lack of adequate frameworks for ethical governance pose serious challenges to ethical data exchange. Guidance on how to best meet these challenges is urgently needed to ensure respect of users' individual rights while ethically harnessing the value of data to spur scientific research, public debate, and societal well-being.

This paper focuses on fundamental ethical issues in data acquisition. In particular, it provides a critical overview of four general data acquisition models, determining and assessing the ethical problems each raises, with a particular focus on protecting people's rights to privacy and autonomy, as well as issues of fairness.

One of the central ethical difficulties facing any data acquisition procedure is balancing the rights of the individuals whom the data is about against the potential benefit to the common good that data offers. Worries about violations of privacy, possible re-identification (even after data has been anonymised), and lack of consent rightly lead to worries about the over-reach of data collectors.1 Such worries, however, should not push us into simply abandoning the project of data collection, as doing so would cripple research (Gymrek et al 2013; Kaye 2012; Mascalzoni et al 2014), and limit our ability to diagnose, treat, and prevent problems (Costa 2014). In particular, collecting large amounts of medical data (i.e., any data about the health condition of an individual) may be especially beneficial in preventing and containing pandemics, diagnosing rare conditions, conducting studies in which effect sizes are small and/or effects occur over long periods of time, and planning medicine production and research.

In light of these worries, we here survey the risks and benefits of four different data acquisition models. The first, opt-in model (Krutzinna et al 2018), poses little threat to privacy and autonomy, but faces problems with generating large, representative data sets, as well as issues about fairness and individual benefit. In contrast, the second, compulsory model, is highly effective at producing a suitable data set, but clearly and egregiously conflicts with rights to privacy and autonomy, and raises distinct problems about fairness.

We next consider the opt-out model (§3) and the market-based model, in which data is effectively purchased (§4). Both of these appear superior to the earlier pair, in that they seem to both be capable of gathering sufficient data whilst assuaging worries about

[1] For further discussion of these particular issues, see e.g. Macnish (this volume) and Vold and Whittlestone (this volume).

privacy, autonomy, and fairness. However, both give rise to more nuanced worries about consent and incentivization. The general upshot is that each of the possible models for data acquisition face important ethical challenges that ought to be taken into account when considering its use.

Before proceeding, a quick note. It is clear that, when it comes to data exchange, specific difficulties will crop up when we consider specific combinations of data donors, types of data, and data collectors—for example, the particular ethical difficulties involved with a large internet company collecting users' internet histories to generate targeted ads will be different than those emerging from a government's collecting citizens' medical data. However, in the following, we make no assumptions about what type of data is being exchanged, or about the individual or agency that is collecting said data. This is done to ensure that the objections we raise are sufficiently general so as to apply to all iterations of data exchange, regardless of the specifics.

OPT-IN MODEL

The first model we will consider is the pure altruistic or opt-in model (Krutzinna et al 2018). The driving idea behind this model is that of particular individuals choosing to 'give' their data (where 'giving' might involve expressly allowing the collection, access, (re)distribution, or retention of said data), without there being any direct or strict incentive for doing so provided by those who collect the data.2 In this way, this model is entirely opt-in, as would-be donors are the final arbitrators of whether their data is collected and used, and donation is driven entirely by something like good will or desire to contribute to the common good.

Such a model governs the exchange of organs in the United States; there, individuals can, if they so wish, give (either pre- or post-mortem) their organs to others. However, there are no direct incentives—the buying and selling of organs is illegal—nor requirements to do so. Such an approach could be readily applied to data exchanges, with would-be data donors giving data acquirers their data (again, pre- or post-mortem), without being motivated by direct incentives or requirements.

What the pure opt-in model does well is protect the donor's rights: because the donor decides to donate their data, worries about lack of consent and, depending upon how we understand the privacy-consent link, privacy look assuaged.3

[2] There is a wide variety of implementations for this model; for present purposes, we assume an idealized, extreme version, in which donors are perfectly aware of how their data will be used. We do so partly for ease of discussion, and partly because this implementation nicely embodies the problems we will shortly highlight.
[3] Of course, this radically simplifies the complex issue of how to go about securing consent, as well as how to understand consent in this context. For further discussion on this, see e.g. Holm & Ploug (2017) and Macnish (this volume). One major issue here is whether individuals get the right to consent to new uses their

However, the model fares badly in several other respects. In particular, because the collected data only comes from those who freely give it, it is unlikely that a sufficiently large/representative data set will be produced. Not only does this dramatically reduce the effectiveness of the data, making it harder to generate wide-ranging conclusions and filter out 'noise', but also brings with it problems about misrepresentation (McNeely and Hahm 2014); for example, if only a small sub-population chooses to donate, then the resulting data will be skewed. This limit can have dire ramifications for practical realizations based on the data, potentially even leading to the misdiagnosis or mistreatment of issues.

Further, the model's entirely opt-in structure encourages free-riding: individuals can readily benefit without having to donate their potentially privacy-compromising data, simply by piggy-backing off the altruistic acts of others. This worry is especially pressing it takes effort to donate data, and/or if donating data carries significant risks with it (such as the risk of re-identification).

Finally, there is a problem about individual benefits. Those donors who allow their data to be collected and used are contributing to a public good and, arguably, they deserve some direct compensation for this.4 However, because it shuns incentivization, the pure opt-in model prevents individuals from specifically benefiting from their individual information capital.

COMPULSORY MODEL

The second model we will consider is the compulsory model. As the name suggests, according to this model, 'donation' is not a matter of choice: an individual's data is simply collected, regardless of their preference on the matter.

While such forced donation may be dismissed as unacceptably totalitarian, it is worth noting that something like this model has been incorporated in democratic states where data collection is tightly integrated into other aspects of social life. In such systems, it is pragmatically impossible to be a member of society without having one's data collected. Denmark's government-run medical data system is one such case, as the collection of medical data is tightly bound-up with tax, election, and other social registries (Havelin et al. 2011; Holm and Ploug 2017), making not 'donating' effectively impossible.

Because the whole population's data is acquired, this model effectively ensures a sufficiently large and representative data set (at least regarding the type of data being

data might be put to after donation; another is whether certain uses of data should, because of the potential for harm, be banned *even if* inidivduals consent to them.

[4] Numerous desert-based views of distributive justice would endorse this claim, provided that data donation is the subject of positive appraisal by others (see e.g. Miller 2001; Mulligan 2018).

collected), which is decidedly a point in the model's favour. But, as with the previous case, there are several points against it.

The biggest problem is that the model runs roughshod over individual rights, bypassing concerns about consent by ignoring donor preferences. The model leaves no room for those who value privacy to avoid having their data gathered and analyzed.

One might respond to this privacy worry by ensuring that the data collected is thoroughly anonymized. Something like this thought seems to underlie the use of Personally Identifiable Information in US privacy law, which, as Schwartz and Solove (2012: 1) note, seems to hinge upon the idea that data that lacks any Personally Identifiable Information poses no genuine threat to privacy.

However, this response faces a significant practical problem, as it is often possible to re-identify people from anonymized data sets. For example, Montjoye et al (2015: 536) studied three months of credit card records for 11 million users. They were able to re-identify 90% of users using only four spatiotemporal points. Similarly, Narayanan and Shmatikov (2008) were able to identify users in Netflix's anonymous database using information gathered from the Internet Movie Database. Finally, Sweeney (2015) was able to match patient names to publicly available anonymized health data.5 These studies show that the claim that anonymized data poses no threat to privacy is dubious at best.

Finally, the compulsory model also faces issues regarding fairness and exploitation. If companies generate profits from using data in research, then they are gaining a valuable resource from individuals without providing any direct compensation.6 This lack of direct compensation is a particular problem if individuals are required to donate their data in order to access necessary services (e.g., medical services), as they are not in a position to (reasonably) refuse—they are, in a way, forced to be exploited.7 Similarly, individuals cannot directly benefit from their particular information capital; their information is simply collected, stored, and used without any kind of payment or direct, specific, individual benefit. In as much as this was a problem for the opt-in model, it is a problem here too.

[5] One way in which worries about re-identificaiton may be alleviated is through better methods of anonymization. For more on issues about re-identification, see Soria-Comas et al. (2015) and Hardy & Maurushat (2017).

[6] This is not to deny that there may be important *indirect* benefits to people, such as a better-functioning medical system. Two points are relevant here. First, the indirect benefits individuals receive may not be enough to address worries about fairness and exploitation. Second, it seems likely that some individuals (for instance, those suffering from rare conditions) will receive more indirect benefits than others—leaving concerns about fairness in place.

[7] It could be objected here the access to the necessary services itself is an important form of indirect compensation. Our response would be that in a just society, access to necessary services should be free of charge.

OPT-OUT MODEL

The two previous models represent two extremes, prioritising the fulfilment of one of the two goals—gathering sufficient, representative data or respecting privacy and autonomy—at the expense of the other. Since neither provides an ideal approach to data acquisition, it is clear that the task is one of finding a suitable middle ground. The two remaining approaches, examined in this and the next sections, are explicitly designed to avoid the problems facing the two extremes.

The first middle ground alternative to the opt-in model and the compulsory model is the opt-out or presumed consent model. On this model, everyone is presumed to give their consent for their data to be collected, stored and used unless they explicitly state otherwise. This presumed consent is taken by some to provide sufficient ethical justification for this practice (Cohen 1992), and is thought to entail an implicit waving of any conflicting privacy or autonomy concerns.

This approach is popular—variants of the opt-out model have been adopted in numerous countries as an appropriate model for organ donation, including France, Wales, and, from 2020 onwards, the Netherlands. Further, it is easy to implement in data exchange contexts: for example, users of a certain phone application might receive a notification informing them that, unless they express an explicit preference otherwise, their data will be collected and used in various ways; the means for 'opting-out' could then be incorporated into one of the application's embedded menus.

As it allows those with concerns about privacy to prevent others from collecting, storing and using their data, the opt-out model has an advantage over the compulsory model. Similarly, because it is likely to lead to much more data being gathered, it also fairs better in that respect than the opt-in model; for example, a study comparing donor and transplant rates in 48 countries over a 13-year period found, as would be expected, that the rates of donation were higher in opt-out versus pure opt-in systems (Shepherd et al. 2014).

However, this model also faces problems. First, there are legitimate worries about how representative the collected data will be. It may well be that certain subgroups of the population disproportionally opt-out—perhaps because they are more informed about problems of re-identification and potential violations of privacy—leading to misrepresentative data sets.

More troubling, however, is that it is not at all clear that everyone who is presumed to consent actually would do so (Veatch and Pitt 1995). In particular, to genuinely consent, it is necessary that the relevant parties be properly informed about the nature of what they are (implicitly) signing up for. But this condition is often not met when it comes to data exchange. This is in part due to the nature of data (it is hard for individuals who are technologically illiterate to understand data collection to the point where they can rightly

be said to give informed consent to their data being collected), but also because how the collected data will be used is constantly evolving. Since how the data will be used in the future is not clear even to the acquiring agent, it is not obvious how data donors can properly consent to said usage. Finally, the mechanisms for opting out are often complex or difficult to discover, effectively forcing many into 'consenting' due to their (literal) inability to say otherwise.[8]

MARKET MODEL

The second middle-ground model is the incentivization or market model. At its heart, this model is driven by individuals opting-in. However, unlike the pure opt-in model, the market model incentivizes donation. This might take the form of some kind of financial incentive (e.g., a straight payment, a tax-break, etc.) or some other direct socio-economic benefit. In exchange for their data, donors receive some kind of payment.

One advantage the market model has over the other three models is that it can lead to Pareto efficiencies. Under normal conditions, we should expect that people will only exchange their data for the incentive if they expect to benefit from this exchange. Similarly, we would only expect a data collector to offer the incentive in exchange for data if they too expect to benefit from the exchange. This model, then, has the potential to make both parties in the exchange directly better off as a result. The capacity of the market-based system to realize Pareto efficiency gains stands in strong contrast to all of the other models we considered so far.

Additionally, under the other three models, the person donating their data receives no direct benefit (of course, they may benefit indirectly from any potential use the data is put to, but there is no guarantee). The market model, on the other hand, provides direct and immediate benefits to the person donating the data. This feature allows those with few resources to benefit from one of the few important economic resources that they may possess: their information capital. The market model allows such individuals to take advantage of this otherwise untapped resource.

The market model also has a number of more specific advantages over the alternatives: A recent overview of various studies investigating the use of incentives to encourage blood donation found that 18 out of the 19 different forms of incentives that have been used were effective in increasing blood donations and that the effects were larger for incentives that were more financially valuable (Lacetera et al., 2013). It is reasonable to think that we would see similar results for financial incentives for data collection, if only because data collection is often less burdensome than blood donation. Hence, the market model is more likely to generate a larger data set than the opt-in model. Further, unlike

[8] Note that this point does not contradict the 'pro' argument concerning easy implementation; that it is difficult to opt-out does not mean that it is difficult to implement the system.

the compulsory model, the market model respects people's right to autonomy, since people can choose not to exchange their data for the financial incentive. Moreover, this model fits with existing practices that are not subject to widespread ethical controversy; supermarket loyalty cards, for example, offer discounts or vouchers to customers in exchange for giving the supermarket permission to collect data about their purchasing behaviour.[9] Similarly, many internet services such as Google and Facebook provide people with a service they can use for free in exchange for the collection and use of their data. We can think of the offering of financial incentives to donate data as simply an extension of these existing uncontroversial practices.[10] Additionally, on the market model, only those who explicitly consent to the collection and use of their data have their data collected. Unlike the opt-out model, the market model cannot be accused of making use of data concerning people who would not properly consent.

Finally, another advantage of the market-based model over the opt-in and opt-out models is that it allows data collectors to target specific groups of the population with personalised incentives. This is useful, as it gives data collectors ways to take action if they suspect their data set is unrepresentative.

For all its advantages, the market model faces several problems. It may, for starters, not generate more data than the opt-in model for personal data that is not so valuable. But even for personal data that would, in principle, be valuable enough to incentivize more people to donate than would do so under the opt-in model, it is difficult to design the market in such a way that it simultaneously protects individuals' rights to privacy and autonomy and generates more data.[11] Many of these disadvantages are instantiations of general worries that have been raised by communitarians and relational egalitarians about markets in goods such as human kidney, surrogacy pregnancy, and sex (Satz 2010; Sandel 2012). We'll discuss three of them here.

First, for the market model to do a significantly better job at respecting individual rights than the compulsory model, there would need to be maximum prices on personal data. Such maximums are needed in order to avoid unduly incentivizing. After all, if the prices for people's data are so high that selling one's data is hard to resist (especially for poorer individuals), then the market would not be respecting people's autonomy.[12] But this is a double-edged sword, because the capped incentive may not be high enough to get people to donate and thus undermine efforts to produce a large and representative data set.

[9] Although most people are likely unaware that this is the deal they are making.
[10] One might argue that the services of these companies offer have become necessary to be part of society nowadays, and people can, consequently, no longer reasonably refuse data collection. If that is so, then the services offered by internet companies should no longer be analyzed from the perspective of the market model, but rather the compulsory model (section 2).
[11] For a longer discussion of these points, see Brouwer *et al* (ms).
[12] We suspect that this worry is especially pressing in the context of medical data.

Second, a market model may lead to a slightly different type of free-riding: there is greater incentive for poorer individuals to share their data, but none for the more wealthy. Consequently, if there are risks associated with donating data, the wealthy are in a position to avoid this risk by not donating while still potentially benefitting from the resulting research. Of course, the free-riding would be less problematic than in the opt-in model, because those who donate their data would receive compensation. However, there would still be potential for free-riding nonetheless.

The free-riding worry is especially pressing given a third concern: such markets may engender exploitation. This concern arises because it is difficult for individuals to establish the value of data, which may make them willing to sell their data for much less than it is actually worth. It may be possible to prevent such exploitation from occurring by heavily regulating the market—for instance by requiring companies to be transparent about all the uses they will put people's data to, imposing minimum prices on data, and requiring them to compensate individuals again if they use data for purposes not initially envisioned. But this is a version of the second problem facing the opt-in model, suggesting that a market approach is in fact no better off in this respect.

These three concerns combine into a fairly demanding constraint on markets in data. It is important to provide an incentive that is both appropriate (i.e., does not lead to undue incentivization) and that will generate a sufficiently large, representative data set.

CONCLUSION

Let us re-assess. We began with one of the fundamental challenges for ethical data exchange: from the perspective of an agent interested in acquiring data, how best do we balance the need to produce a sufficiently large, representative data set (and all the actions that generating this requires) with the rights to privacy and autonomy of the individuals the data is about? Taking this challenge seriously, we then explored two extreme models, each of which maximizes one particular element: the opt-in model, which goes all in for privacy and autonomy, and the compulsory model, which guarantees a good data set. Both of these models were shown to face substantive objections, both ethical and practical. If we were to solve the challenge, it was likely to not be with one of these two extreme models.

This result motivated exploring two more moderate approaches. The first of these was the opt-out model, which turns upon a notion of presumed consent. While this approach seemed better than both the opt-in and the compulsory models, it too faced difficulties. The biggest problem concerned ensuring that data donors were genuinely consenting; otherwise, this model in fact turns out to be just as ethically problematic as the compulsory model. The second was the market model, which takes the basics of the opt-in approach, but adds direct incentives to motivate would-be data donors. Such an

approach has numerous advantages over the other three. However, an ethical market requires a cap on the incentive provided. Too low an incentive may be exploitative, and may also undermine the ability of the market approach to do better than the opt-in model when it comes to producing a large, representative data set. Meanwhile, too high leads to undue inducement (and hence undermines the donors' autonomy). There does not seem to be any clear way to find the happy medium—the choices force a compromise, or trade-off. The market model too faces substantive ethical objections, but it seems to be the one with a greater chance to strike a balance between respect for rights and gaining access to high-quality data.

The general upshot is that all four of these approaches lead to distinct ethical difficulties. However, before concluding, we would like to briefly mention one other option: so-called blended models (see e.g. Poikola et al 2014). These involve blending together elements from various other theories and approaches—for example, one might combine the compulsory model for exchanges of 'basic' or non-sensitive data, together with an opt-in model for more sensitive data (e.g., medical or financial data). Blended models hold out the promise of delivering all the advantages and none of the disadvantages of the components. However, we are not very optimistic. We suspect that modified versions of the general objections raised above will apply to the relevant parts of these blended models—i.e., using the example blended model above, it will still be the case that privacy concerns are going unheeded with regards to 'non-sensitive' data; similarly, the approach is unlikely to generate a sufficiently large, representative pool of 'sensitive' data. It is possible that some clever combination would be able to avoid these objections. Yet no such model presently exists. We thus conclude that those considering the ethics of data acquisition must confront the implicit dilemma of data collection and engage in inescapable ethical trade-offs. That is, data ethicists ought to focus their efforts on how to comparatively weight the value of consent, privacy, or fairness against the proposed social benefits of data collection. The result of such comparisons may differ in specific cases, but cannot, as we have illustrated, be resolved in general via particular models of data acquisition. Although the market model seems to be the most promising framework to reach a satisfactory balance between conflicting goals, it is of utmost importance to assess each case on its own merit in order to establish how to best protect rights, what is an appropriate risk acceptance level, and what counts as fair compensation.

REFERENCES

Brouwer, H., Cawston, A., Wildman, N., Archer, A., and Bradley, S. (*ms*). Dollars for Data? Critically assessing the market model for data exchange. Unpublished manuscript.

Chretien J-P, Rivers CM, Johansson MA. 2016. Make Data Sharing Routine to Prepare for Public Health Emergencies. *PLoS Med* 13(8): e1002109. https://doi.org/10.1371/journal.pmed.1002109.

Cohen, C. 1992. The case for presumed consent to transplant human organs after death. *Transplantation Proceedings*, 24: 2168–2172.

Costa, F. F. 2014. Big data in biomedicine. *Drug Discovery Today* 19(4): 433–440.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339(6117): 321–324.

Hardy, K. and A. Maurushat. 2017. Opening up government data for Big Data analysis and public benefit. *Computer Law & Security Review* 33(1): 30-37.

Holm, S. and T. Ploug. 2017. Big data and health research – the governance challenges in a mixed data economy. *Bioethical Inquiry* 14: 515-525.

Kaye, J. 2012. The tension between data sharing and the protection of privacy in genomics research. *Annual Review of Genomics and Human Genetics* 13(1): 415–431.

Krutzinna, Jenny and Taddeo, Mariarosaria and Floridi, Luciano, Enabling Posthumous Medical Data Donation: A Plea for the Ethical Utilisation of Personal Health Data (April 1, 2018). http://dx.doi.org/10.2139/ssrn.3177989.

Lacetera, N., M. Macis, and R. Slonim. 2013. Economic rewards to motivate blood donations. *Science* 340 (6135): 927-928.

Mascalzoni, D., A. Paradiso and M. Hansson. 2014. Rare disease research: Breaking the privacy barrier. *Applied and Translational Genomics*, 3(2): 23–29.

McNeely, C. L., and J. Hahm. 2014. The Big (Data) Bang: Policy, prospects, and challenges. *Review of Policy Research* 31(4): 304–310.

Miller, D. 2001. *Principles of social justice.* Cambridge, MA: Harvard University Press.

Montjoye, A., L. Radaelli, V. K. Singh, and A. Pentland. 2015. Unique in the shopping mall: on the reidentifiability of credit card. *Science* 347(6221): 536-539.

Mulligan, T. 2018. *Justice and the meritocratic state.* New York: Routledge.

Narayanan, A., and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY: 111-125.

Poikola, A, Kuikkaniemi, K. and Kuittinen, O. 2014. My data – johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen. Liikenne- ja viestintäministeriö. http://www.lvm.fi/julkaisu/4420389/my-datajohdatus-ihmiskeskeiseen-henkilotiedon-hyodyntamiseen

Sandel, M. J. 2012. *What money can't buy: the moral limits of markets*. (New York: Farrar, Straus and Giroux).

Satz, D. 2010. *Why some things should not be for sale: The moral limits of markets*. (Oxford: Oxford University Press).

Schwartz, P. M., and D.J. Solove. 2011. Pii 2.0: Privacy and a new approach to personal information. *Privacy and Security Law Report.*

Shepherd, L., R.E. O'Carroll and E. Ferguson. 2014. An international comparison of deceased and living organ donation/transplant rates in opt-in and opt-out systems: a panel study. *BMC medicine 12*(1): 131.

Soria-Comas, J., J. Domingo-Ferrer, D. Sánchez, and Sergio Martínez. 2015. t-Closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering* 27(11): 3098-3110.

Sweeney, L. 2015. Only you, your doctor, and many others may know. *Technology Science 2015092903*.

Veatch, R. M., and J. B. Pitt. 1995. The myth of presumed consent: ethical problems in new organ procurement strategies. *Transplantation Proceedings* 27(2): 1888-1892.