EasyChair Preprint

№ 1059

Nudging the sharing of data in a fair way

Michele Loi, Matteo Galletti and Paul-Olivier Dehaye

May 28, 2019

Nudging the sharing of data in a fair way

Introduction

One aspect of the digital transformation of commerce and social relations is the enhanced capacity of digital infrastructure to influence individual choices. As this happens, legal contracts become progressively less relevant causes of online choices at least with respect to personal data, as some privacy scholars have recognized (Hartzog 2018). In other words, the shape or color of an action button, or the time needed to access and select data sharing preferences, tend to be more influential causes of data protection behavior than the textual content of terms and conditions, which few people read. Privacy is not, however, the only value at stake when individuals and businesses make choices online concerning their personal data. Justice is another such value. The importance of data for justice has not gone unrecognized. We might distinguish three types of discourses around the justice of data: data used for governance can undermine justice by enhancing power asymmetries (Johnson, 2014), data technologies can enhance justice by making the poor visible (Heeks and Renken, 2016) and justice can be said to obtain or not obtain in the distribution of economic benefits of the data economy (Mann 2016).

None of these discourses, however, provides direct guidance concerning the question whether digital nudges are just. By nudges we mean the decisions of choice architects, people who create environments in which other people make choices. Nudges are obviously of great importance for online interactions, which take place in constructed virtual environments. Different views have been expressed about the morality of nudges, ranging from an eager acceptance of such tools (Nys and Engelen 2017; Schmidt 2017) to a moral criticism of them because of their manipulative nature (Bovens 2009; Godwin 2012). One of the most important ethical framework of assessing nudges, by some of the leading scholars in this debate (Sunstein 2014) involves four values, curiously not including justice. In principle Sunstein's frameworks aims to protect individuals from harms and provide them with goods. We could resume the core of his arguments with this sentence: "the goal of many nudges is to make life simpler, safer, or easier for people to navigate" (Sunstein 2014, 584). Sunstein claims that

some nudges don't benefit individual choosers, but are designed to promote fairness, by invoking social norms or default rules that favour collective disposition of charity and generosity. When nudging produces an unfair condition (e.g. discrimination), the remedies he considers to counteract that side-effect are educative campaigns and legislative intervention, but he doesn't offer practical and ethical indications to build fair choice architectures (Sunstein 2016, 34-35).

Yet the question of the justice of nudges is a natural one to ask. More specifically, it is an important question for data justice, as characterized by Linnet Taylor (Taylor 2017). As our ethical analysis shall show, nudges affect all three dimensions of data justice Taylor identifies, namely visibility, which concerns the informational privacy of the individual, digital (dis)engagement, which considers the ability of the individual to both benefit from data and obtain benefits independent of them, and data-driven discrimination, which concerns the biases and (implicitly or explicitly) discriminatory choices taken on the basis of data. For ease of exposition, and for practical relevance, we focus our analysis on online digital platforms, which are arguably the sole uniform digital environments currently able to affect the lives of multitudes. We also focus on private companies, as public administrations are reasonably expected to satisfy higher standards of justification compared to entities with which citizens voluntarily choose to engage.

Here we propose a theory to fill this gap. Our approach can be summarized as follows. We distinguish educative from non-educative nudges, namely, those features of choice environments that engage with the user's rationality (educative) and those that engage with his heuristics and biases (non-educative). We argue that non-educative nudges constitute a power – the power to exert influence on the choices of another human being – that is necessarily unequally distributed between choice architects and platform users. We assume that a power inequality is unjust, unless it is justified. Hence, we provide three principles of platform justice that justify non-educative nudges. Our three principles are loosely inspired by Rawls's two principles of justice and are justified by relying on the Rawlsian methodology of reflective equilibrium. More precisely, we distinguish six dimensions of nudging power exercised by online platforms and analyze, through examples, the implications of these three principles concerning each of them. We maintain that their implications are plausible, as they indicate a sophisticated, yet principled, middle way between the opposite extremes of considering all non-educative digital nudges morally impermissible in themselves, and of considering all nudges morally unproblematic.

The phenomena to which justice can be attributed differ in terms of their levels of abstraction. Here we consider the level of the privately owned online platform as a voluntary association between individuals, and as a free association in liberal-democratic society in Rawlsian jargon (Rawls 1999, 412–13). The principles of platform justice we offer are ideas for a possible morality of this peculiar kind of association (Rawls 1999, 412–13), which is compatible with the business ethics idea that even business, especially large, globally influential ones, have moral obligations (O'Neill 2000). It is also compatible with the idea of business as a social contract (Donaldson and Dunfee 1999) evaluated with standards drawn from actual moral practice but critically evaluated in the light of hypernorms, widely recognized higher norms by which all other norms are judged. The third principle provides a justification for many nudges which considered uncontroversial and used within platforms, while the first and the second reflect hypernorms concerning individual security and non-discrimination.

1. Nudges and choice architectures

As the subtitle of a book by Cass R. Sunstein emphasizes, we're living in the age of behavioural science. In recent decades, studies in social and cognitive psychology mapped the cognitive and volitional potentialities and limits of human beings, providing a huge amount of empirical data. "Behavioural informed approaches" in policy-making try to capitalise on this gold mine to improve policy and create environments where individuals can make better decisions. The "nudge approach" is in line with this trend, since it tries to steer people's behaviour towards goals relying on empirical knowledge on how the mind works. Altering relevant environmental features (for example, framing information in a certain way or making some options more salient), this approach influences individuals to choose the better option, all the while retaining freedom of choice, because they're always free to choose and act otherwise. Public and private decision-makers are therefore "architects of choice" who have "the responsibility for organizing the context in which people make decisions" (Thaler and Sunstein 2011, 3) and the task to help people to make good choices in a prudential or moral sense. Better prudential choices have a positive impact on the wellbeing of the individual chooser; better moral choices have a positive impact on a common good, such as a limitation of pollution, or on maximizing moral behaviour, such as altruistic actions. Thaler and Sunstein's formulation of the nudge approach is the best-known and most influential one. According to their definition, a nudge is "[...] any aspect of the choice architecture that alters people's behaviour in a predictable way, without forbidding any options or significantly changing their economic incentives" (Thaler and Sunstein 2011, 6).

This definition, however, is too wide, because it implies that the alteration of disincentives can be counted as a "nudge." As Hansen and Jespersen argue, this definition does not exclude from the nudges set the influence of behaviour via infliction of pain; so architects could associate choosing a certain option with receiving a shock of 10,000 volts and this choice architecture should be enlisted among the nudge interventions, because it does not increase the incentive of any options (Hansen and Jespersen 2013, 6). To avoid this paradoxical consequence, they suggest integrating the original definition with the one introduced by Hausman and Welch, so as to include incentives and disincentives in the list of proscribed interventions on people's choice set: "Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth" (Hausman and Welch 2010, 126). So, we can define nudges as ways of influencing people's choices predictably without limiting the choice set, or significantly changing their economic incentives, or making alternatives appreciably more costly in terms of money, time, trouble, social sanctions, and so forth.

It should be noted that, while capturing some peculiar aspects of choice architecture, the amended definition, as the original one, covers by itself a wide-range catalogue of influences. Beside interventions that make use of individual or collective biases and cognitive flaws, Sunstein introduces what he calls "educative nudges" which "attempt to inform people, or to build their competence, so that they can make better choices for themselves [...] to overcome or correct behavioural biases by promoting learning – or more modestly, by creating a choice architecture in which those biases will not manifest themselves" (Sunstein 2016, 32; see also Sunstein 2012, 59-60). By "educative nudges" we mean those that neither coerce individuals, nor alter the incentives or cost of the options, nor exploit cognitive flaws, but they

correct or enhance understanding, reasoning, or decision-making competence.[1] Actually, educative nudges rely on research programs in psychology other than the “heuristic and biases” program on which the nudge approach is based and so it’s not clear if they can easily be grouped with non-educative interventions (Hertwig and Grüne-Yanoff 2017; Reijula S. et al. 2018).

However, in general the nudge approach profits from a partition of psychological processes in two systems with different operational features (Thaler and Sunstein 2011). The processes that make up the so-called "System 1" are fast, parallel, automatic, associative, do not require efforts; those that make up the so-called "System 2" are slow, in series, controlled, require efforts and are governed by rules. A first group of nudges exploits the processes of the system 1 which, in many situations, cause non-optimal responses to environmental challenges.

The following list is not comprehensive, but exemplifies the main System 1 nudges:

Default rules.

Many studies have shown that we generally incline to show inertia and status quo bias: we tend to stick to default rules without making choices that can change them (even when status quo is not beneficial for us). So, the choice architects may set the most rational or beneficial option on the knowledge that people are likely to not change it, even if they’re free to do so. For example, in an office environment printers can be set to print on the both sides of the sheets so to reduce paper use, because those who use the machines will tend to not change the set condition; in welfare systems that allow it, employees are automatically enrolled in the retirement plan most beneficial for them. They can opt out and choose another plan, but they hardly do so because of inertia bias.

Uses of social norms.

Emphasizing what most people do can obtain relevant effects on individual behaviour, because anyone is inclined to conform to what her social group does. (But on the complex relations between nudges, social norms, and expectations see Bicchieri and Dimant 2019.)

Framing effect.

People tend to react differently to the same information framed in different way. For example, disclosing the survival rates of patients treated with a therapy increases the probability that patients accept it, where disclosing the mortality rates has the opposite consequence.

Warnings, graphic or otherwise.

To augment the salience of an option or information, choice architects can use various gimmicks, as using bright colours or inserting graphic pictures (see the images of harms produced by smoking on cigarette packets).

Educative nudges target System 2 and try to enhance our cognitive or volitional processes. For example, Sunstein and Thaler quotes the site stickk.com, where individuals can entrust a certain amount of money they can lose if they can’t reach a certain goal, under the supervision of a trusted person. This could a good way to enhance self-control in losing weight or quitting smoking, because the potential loss of money motivates agents to honour their commitments (Thaler and Sunstein 2011, 231-2). In that and in other cases, the architecture of choice does not use or exploit our cognitive or decision limits but try to overcome them, increasing the capacity for self-control and understanding.

[1] Partial information is not an educative nudge when the information is misleading. In order to count as educative, the information must *correct* or *enhance* understanding and decision-making capacity, not enhance *misunderstanding* and *undermine* decision-making capacity.

Intuitively non-educative nudges are more problematic than educative ones, because they exploit automatic cognitive shortcuts and defects, bypassing deliberative and conscious processes, to produce the intended behaviour. Unlikely, educative nudges can bring about a certain behavioural outcome, but their primary target is competence; in principle, they don't bypass deliberative capacities. Most times, as in the stickk.com example, the agent chooses them freely, as aids allowing her to circumvent her weaknesses and to achieve significative goals. Sunstein himself acknowledges that some interventions are more problematic than others and that, for example, warnings, labels on products and services that specify the calorie intake or energy conservation influence individual behaviour without manipulating it (Thaler and Sunstein 2011, 80).

Obviously, such an evaluative judgement depends about the values at stake that nudges are supposed to erode. In the next sections we discuss the relationship between autonomy and welfare to show that to morally evaluate nudges we should adopt a broader view on autonomy than Sunstein's.

2. Online nudges and choice architectures

In this section we provide a catalogue of different actions that may be nudged in the context of internet platforms. The topic of nudging is established in the literature on human-machine interaction (Weinmann, Schneider, and Brocke 2016). Digital nudging has been defined as

the use of user-interface design elements to guide people's behavior in digital choice environments [such as] user interfaces – such as web-based forms and ERP screens – that require people to make judgments or decisions (Weinmann, Schneider, and Brocke 2016, 433).

Examples of user interface design elements are pop-up windows asking the web-site user if he wants to receive notifications, and the 'yes' button has a brighter color than the 'no' button (salience). Other examples are websites designed to display the running cost of a service (incentives), changing defaults in the consent for data collection (defaults), symbolic rewards and goals in gamification (feedback), requiring people to type their password twice (expecting errors), and guiding users through the purchase process of a complex product (structure complex choices) (Weinmann, Schneider, and Brocke 2016) .

Here we focus on the user-interface design of online platforms, which we define as an online digital environment allowing peer-to-peer or peer-to-provider communication or commercial exchange, where by 'provider' one means a service provider other than the platform provider itself. This excludes ordinary commercial websites allowing only communication and commercial exchange between the web-site user and the service provider; it does not exclude those platforms, like the Amazon website, that enable both communication and commercial exchange between the user and the platform owner and between the user and other service providers (e.g. independent sellers on the Amazon platform). Furthermore, our definition is restricted to privately owned platforms, but it is not restricted to platforms pursuing primarily commercial goals *between peers*. Non-primarily commercial platforms, such as search engines (e.g. Google), social networks (e.g. Facebook), microblogging sites (e.g. Twitter), and political deliberation and voting platforms (Rousseau, the privately owned platform service used by the Italian Movimento Cinque Stelle) are also included. Public platforms, e.g. platforms for the provision of services by the public

administration, for enabling e-voting and public consultations, are excluded because they raise peculiar problems, having to do with special requirements of transparency and accountability to the public. Moreover, the scope of our paper is limited to platforms participation to which is *voluntary*. This excludes private platform which individuals are coerced to avoid.

The catalogue categorizes the actions nudged by platforms in different dimensions. All Dimensions regard choices that affect the distribution of resources or the welfare of the chooser. Dimensions differ because different dimensions concern different types of choices. A type of choice corresponds to a resource directly controlled by the choice.

Dimension 1: Privacy

This dimension concerns the choice of *reviewing and choosing privacy options and data shared with and through the platform*. The resource directly controlled by the choice are the data co-produced by platforms and platform users (usually referred to as *user data collected by the platform*, but we take a different view of this data than most). These data may be personal, anonymous, or anonymized, produced by the interaction of platform software and platform user behavior.

Some kind of nudge is virtually unavoidable as privacy and data sharing choices are standardly set online. Intentionally or unintentionally, the design of web pages and procedures, necessary to express the informed consent will always affect the way privacy and data protection preferences are expressed. For example, a platform can select sharing or not sharing specific kinds of data, with the platform and/or its partners, as the default setting;[2] furthermore, it can encourage or discourage access to the settings page by making it easier or harder to land to that page, or by increasing or decreasing the perceptual saliency in the way certain options are visualized (e.g. text size, color, etc.).

Nudging privacy and data sharing properties has an influence on the interests of platform users and owners. For example, different amounts of sharing data with the platform affect the degree to which:

- some or all individuals are exposed to heightened risk that their private information may be revealed if the platform is hacked;
- some individuals are offered higher/lower prices than others (e.g. in dynamic personalized pricing);
- individuals are exposed to the influence of the platform, affecting the risk of entrapment (addiction to the service/polarization), and to the desires of other users, including advertisers, affecting political preferences among others.

Concerning the benefits for platforms:

- the platform owner is able to capitalize data assets from which it benefits via:
  - data commercialization,
  - providing other users more finely stratified access to the users.

User addiction/entrapment, engagement, data commercialization, and the ability to offer stratified access to the users to third parties or other users all contribute to strengthen the platform by enhancing its long-term profitability. Often, long-term profitability is achieved indirectly, via the creation of market dominance, cemented through network effects that grow with the amount of users in the platform. User data, enabling stratified access to the users, are the main capital for achieving profitability. First of all, they are crucial capital directly, when stratified access is an aspect of commercial services for paying users or third parties (e.g.

[2] We abstract away from legal requirements, which differ in different jurisdictions. The EU General Data Protection Regulation for example has specific requirements about privacy defaults.

targeted advertising, for advertisers). Secondly, they are capital indirectly: by improving the services offered to all users (e.g. the ability to find a friend, or a relevant social group, on a social network), they make the platform more attractive to them, which helps the platform to grow larger and become more dominant.

Dimension 2: Beneficiaries

Users in a platform typically select other users or third parties as partners for special interactions. In social networks, users select private persons, but also companies or political personalities, as privileged partners in communication and *exchange*. In online marketplaces like Amazon, Ebay, and Booking.com, in peer production and sharing economy platforms e.g. AirBnB, Uber, in content platforms (Netflix or Apple Music), in all platform subsidized by advertising (e.g. Google Search) users select providers of goods or services, as their primary goal for using the platform or following exposure to advertising, or both (YouTube, Facebook, Google Search, Amazon, Ebay, Google). In online donation platforms, such as DonorsChoose, Chuffed, GlobalGiving, CrowdFunder or Facebook fundraising, choice architectures shape donors giving behavior without shaping their choices, for example "the default ranking strategy used in project listings, decision making aids such as search filters, or the selection of project attributes disclosed to the donors" (Chakraborty et al. 2019).

In the first example, the special interaction partners are beneficiaries of goods of the platform users: in the first case of the good attention, in the second, of the good of opportunities for economic transactions, in the third, of charitable donations. Users can be nudged to purchase goods from companies via their position in the ranking of search results, where the ranking is based on pertinence, advertising investment by the company, or combination of both criteria. Choice architectures may direct philanthropic spending through rankings based on criteria like location, type of previously funded projects, type of donors, etc.

Notice that such nudging can be in the interest of users, not only platforms, as it may be conducive to higher user satisfaction. Or at least, this seems apparent if overall user satisfaction is measured in an utilitarian way as the aggregation of signals of immediate satisfactions, such as aggregate click-rates, amount of purchases on platforms, amount of charity donations. These utilitarian user satisfaction metrics are entirely blind with respect to the question of the fairness of such choices; they could be critically re-evaluated and lead to a lower level of satisfactions in reflective users. Users, that is, may be less satisfied when they realize that the same choice architectures may be co-responsible of generating inequalities. A case in point is that of crowdfunded charity platforms. As it has been pointed out "The core mission of such platforms is to counter existing social inequalities. Biased donor behavior, however, may skew the distribution of donations towards or against certain recipients, exacerbating some of these inequalities as a result" (Chakraborty et al. 2019).

Dimension 3: Cybersecurity

It has been observed that adopting cybersecurity precautions reflects a utility-optimizing behavior that is sensitive to both the individual actual costs associated with the behavior (login speed, the value of information in the online account) and to subjectively estimated probability of harm (protection offered by a behavior, risk of hacking) (Redmiles, Mazurek, and Dickerson 2018). It is therefore not surprising that cybersecurity behavior can be influenced by affecting both the perception of costs and of risk through choice architectures (Briggs, Jeske, and Coventry 2017). Fig.1 below presents a framework of nudges for cybersecurity, developed by applying the MINDSPACE (Dolan et al. 2012) choice architecture framework.

Influencers	Description of possible nudges (for chosen scenario)
Messenger	Warning messages should come from a trusted provider [...]
Incentives	When connected to an unsecure network hamper productivity [...]
Norms	Tell the users the % of people who lost/infected data within the company that have used that network [...]
Defaults	Present most secure as first option [...]
Salience	Prompt 'not a secure network' [...]
Affect	Use emotive colors [...]

Table 1. Nudges for cybersecurity. From (Coventry et al. 2014)

Cybersecurity appears one of the dimensions in which the use of nudges is easier to justify in so far as users have to balance short-term certain gains (higher usability, typically deriving from lower login costs) with long-term, merely possible harms (probability of being hacked). The benefits for users and platforms also appear especially aligned: both platforms and users stand to gain from higher security in the long term; both stand to lose from heightened protections in the short term when these are achieved at the expense of usability.

Dimension 4: horizontal expansion of services

By 'horizontal expansion of services' we mean nudges on users to install services delivered through specific providers.. For example, users of iOS (the mobile operating system by Apple) are nudged to use the financial service provided by Apple, Apple Pay, through constant push notifications encouraging them to set it up. A clear benefit for the platform provider is that it makes it easy for the user to explore new types of services provided by the same provider. While, being a nudge, users are not coerced to use the Apple service, still the choice architecture cements dominance by (non-coercively) affecting the behavior of the user. This may eventually harm consumers if competition is reduced, limiting the choice for other products and the incentive for a dominant platform to innovate. It might even tip entire business ecosystems, as upstarts know their new products will have to conform to the dominant platform's choice architecture in order to have a chance to compete (for instance, the news industry is struggling financially because of the dominance of social media platforms as distribution channels, killing both classified ads and subscription news as viable business models, and limiting their agency in deciding how to address the situation).

Dimension 5: politics

The dimension of politics, as we defined it, consists in the fact that users may be nudged by platforms to address their political representatives. The most widely known example is Facebook's "get-out-the-vote" initiative, which demonstrated its ability to nudge people to vote, using a "your friend voted" social nudge (Bond et al. 2012). Such power can affect the electoral turnout by hundreds of thousands of votes in a country like the US. Such nudges confer disproportionate power to platform owners but they can also be used to promote worthy causes (e.g. it could be claimed that getting more people to vote, irrespective of political affiliation, is a worthy, politically neutral, democratic cause). Another, significantly

more controversial, example, is Uber using its app to direct its customers on a petition site, to sign against the New York City major decision to freeze all new licenses of for hire vehicles (Bohn 2015). A further, more complex case, is that of a private platform supporting deliberation and e-voting by a private association – e.g. a political party – which exerts political power by electing representatives to Parliaments. A case in point is the Rousseau platform used by the Italian political movement (and self-described 'non-party') Movimento Cinque Stelle. A political deliberation platform is a choice architecture when platform owners/managers decide, for example, the formulation of the political questions that users can vote on, the voting procedure, the mechanism that enables users to define a political agenda, fully bottom-up, or provide feedback and/or complete in an interactive way the contents proposed from the platform management. All these choice architectures, beside the user-interface design, bear on the probability that certain options will be voted and that certain themes, rather than others, will be discussed on the platform.

Dimension 6: policing of social norms

Last but not least, platforms can use their choice architectures in order to nudge the policing of social norms, by which we mean reporting the behavior of other users incompatible with terms and conditions. Choice architectures may be used to make users differently likely to report such violations. A significant benefit for companies is that in this way they acquire a significant power to influence the environment shared of their user.

The main benefit from platforms is control over the environment of user interaction. This is extremely important for commercial platforms because an uncontrolled environment can easily degrade into a toxic one, discouraging user engagement and retention. Moreover, users also benefit from a reasonably polished environment, especially those users belonging to minorities that are more likely to be exposed to harm in the form of hate speech. The amount of user engagement and retention is a simple utilitarian measure of user well-being (insensitive to distributive issues and to reflexive endorsement of behaviorally revealed preferences). In this sense, the platform owners' interests and those of users are fully aligned.

In theory, however, control over choice architectures gives platform owners the power to influence the enjoyment of freedom in a politically partisan way. That would clearly give platform owners great political power. It is, however, reasonable to assume that most private platform owners will not want to exercise power in such an explicit form. The reputational risks are too high, as most individuals do not like to be nudged in a politically partisan direction. For that reason, platforms are more likely to intentionally design choice architectures that affect the probability of policing content without pursuing partisan political goals, at least directly. However, nudging on terms and conditions policing is never politically neutral even if exclusively non-politically-partisan goals are pursued, for at least two reasons. First, because the question of the balance between freedom of speech and harmful speech is a substantive political question, as we shall discuss in section 5. Second, because even procedurally neutral criteria may have a disparate impact on political views that are differently disposed to be expressed through borderline harmful content.

Dimension	Action nudged
1. Privacy	Reviewing and choosing data sharing and privacy options
2. Beneficiaries	Selecting of beneficiaries of users' interactions
3. Cybersecurity	Adopting cybersecurity precautions

4. Horizontal expansion of services	Installing services delivered through specific providers
5. Politics	Addressing elected representatives
6. Policing of social norms	Reporting behavior incompatible with terms and conditions

Table 2. Dimensions on nudging on platforms

3. The ethics of nudging: autonomy

One way to evaluate the moral acceptability of nudging is to use the grid of values Sunstein introduces in the chapter 4 of The Ethics of Influence. There he analyzes the impact of nudging on four basic values: welfare, autonomy, dignity, and self-government. We can't discuss all Sunstein's argument in this paper, so we focus on the role of autonomy in the moral evaluation of nudging.

Autonomy is a relevant value because online nudging, as other similar choice architectures, can exert a manipulative influence on users. On the one hand, Sunstein wants to stay in the wake of the liberal and Millian tradition (although with an updated empirical background) and he should be concerned with the effects on personal autonomy (on the compatibilism between the Millian liberalism and the nudge program see also Brink 2013, 194). On the other hand, in the current debate, one recurrent objection to nudging is that it has a manipulative nature and a negative impact on nudgee's personal autonomy. Even if nudges are not-coercive means of influencing human behavior, they have a distortive effect on deliberation and choice. The philosophical literature about the conceptual and ethical nature of manipulation is growing fast and we can't re recall the entire debate here on manipulation; one prominent theory claims that the wrongness of manipulation is relative to its consequences on personal autonomy. So, it's fundamental to have a detailed account of personal autonomy to evaluate the actual effects nudges have on it.

Generally, Sunstein claims that the wrongness of manipulation depends on the fact that this means of influencing choices bypasses or insufficiently engages the manipulee's capacity for reflective and deliberative choice (Sunstein 2006, 82-90; for other definitions of manipulation see Coons and Weber 2014). According to this definition, some nudges seem to be morally wrong because they don't involve rational deliberation. For example, default rules exploit the inertia bias and bypass reflective deliberation. Nevertheless, Sunstein argues that "life cannot be navigated without default rules, and so long as the [choice architect] is not hiding or suppressing anything (and is thus respecting transparency), the choice of one or another should not be characterized as manipulative" (Sunstein 2006, 92). Educative nudges are clearly non-manipulative, because they enhance understanding and autonomy. The problem is that infringements on autonomy constitute a broader class than acts of manipulation. To see it, we should contemplate a richer map of the dimensions of autonomy.

Sunstein introduces two ways of understanding autonomy. According to the first way, autonomy can be deemed a feature allowing individuals to control their desires and decisions, according to their values and ideals. People not only have desires about certain things, they

have also preferences about those desires, so that they can adopt a reflective standpoint to be guided by desires (or pro-attitudes) they can endorse. Choice architects, Sunstein claims, should follow people's second ordered critical judgments produced by System 2 (Sunstein 2006, 49). But later Sunstein suggests another analysis of autonomy, that reduces it to freedom of choice conceived in an "economic" way: what is relevant is not the role of the choice in the mental economy of the chooser (a choice is autonomous not if the agent has a positive preference about the desire that motivates it), but whether there are external costs imposed on choosers (Sunstein 2006, 63-64). Understanding autonomy in this minimal way, Sunstein can easily argue that nudges don't infringe on autonomy because a condition to identify a choice architecture as a nudge is that it has not to impose excessive costs or burden (in terms of incentives or disincentives) on choice.

Educative nudges clearly respect this condition, because information, (graphic) warnings, and reminders do not seem to create a risk to freedom, not more than a friend who tries to convince you to stop smoking, to move to another city, or to leave your job (Sunstein 2006, 64). Non-educative nudges don't affect autonomy (defined minimalistically), either, because they are "designed to ensure that choices are informed and that relevant information is salient and easy to process" (Sunstein 2006, 65). Default rules can pose specific problems because agents failing to change the default rule for the power of inertia don't act for their own reasons. Choices are autonomous in so far as agent exercise her authority over them and the reasons that motivates them, but, in the case of default rules, actions stem out from reasons that don't flow from the agent's authoritative power, but from a bias. Sunstein rejects this picture of autonomous choice because it's too demanding. It is untrue that agents act always on reasons, especially when the stakes are low, and sometimes they act on reasons that are not consistent or well-integrated into their evaluative standards. In some areas of life (mobile phones, printer settings, mortgage, rental contracts, energy and savings plans), default rules can be warranted (Sunstein 2006, 67).

We can understand the double meaning of autonomy as introduced by Sunstein as an interesting way to capture the intricacy of the concept of autonomy, but some authors emphasized that it's not enough. For example, Blumenthal-Barby broke up the concept of autonomy, identifying different components. They are the classical items listed in Faden and Beauchamp's definition of autonomy: (1) understanding and appreciation; (2) intentionality; (3) absence of controlling and/or alienating influence (Blumenthal-Barby 2016, 7), to which she adds another important component described by Schwab (2006), effective autonomy. Effective autonomy "is the matching of formally autonomous interests or desires with decisions that will achieve those interests or desires" (Schwab 2006, 575).[3]

We have then a useful map to evaluate whether single nudges can impair one or more levels of autonomy, affecting the individual capacity of self-ruling. So, even if nudges are not manipulative in Sunstein's sense, because they don't bypass or fail to engage sufficiently deliberative and rational capacities, they can negatively affect one or more dimensions of personal autonomy. Sunstein seems to claim that there are other reasons that can outweigh the loss of autonomy in certain cases and they're welfare grounded reasons; as manipulation, sometimes a choice architect can nudge an individual to make a non-autonomous choice with

[3] There's further level that we don't discuss and consider in judging the use of nudges on digital platforms; it's what can be called "authenticity" level. For authenticity we mean here the general coherence between the individual's actual choices and the pattern of cares and pro-attitudes toward values, projects, relations that she deems important in her own lives. There's a general resistance in literature to take seriously the whole impact of non-educative nudge on this dimension of autonomy (see for example, Gorin et al. 2017 and Holm 2017).

the end of enhancing the nudgee's autonomy on the long run. A provisional conclusion can be that non-educative nudging is pro-tanto wrong and can be morally permissible, or even required, all things considered. One especially important case of justified non-educative nudge is one that enhances the subject's welfare coherently with system 2 judgment and with the reflective preferences of the nudged subject.

4. Data justice and platform nudges

4.1. Philosophical theories of justice: how they differ

Justice concerns the distribution of goods among claimants. Philosophers disagree concerning the kind of good that expresses the best metrics of whatever it is that should be distributed justly: the currency of justice. Resources, welfare, opportunities for welfare, capabilities, have all been mentioned as relevant in the literature.

Furthermore, an important distinction is between end-state and historical principles of justice. End-state principles define a distribution of goods among individuals as just based on a structural property of a distribution, ignoring how that distribution came about. One example is simple resource egalitarianism. Justice is achieved if everyone has equal resources. As an instance of this, the 'instrumental account' of data justice described by (Heeks and Renken, 2016) characterizes data justice in terms of the desired end-states from collecting and using data. The relevant end-states concern the distribution of broad social goods, e.g. individuals' capabilities.

Historical principles define a distribution of goods as just based solely on information concerning how the distribution came about. For example, a simplified version of Robert Nozick's libertarian theory defines a just distribution as one that results from voluntary exchanges between adults of legitimately owned resources, where any entity can be legitimately owned if it does not have an owner, yet (Nozick 1974). We shall refer to such theories as 'procedural-historical'. In the data-justice debate, Heeks and Renken (2016) characterize these as procedural. In this account data justice results from the fair handling of data, and 'fair handling' is defined by purely deontological-procedural notions such as informed consent, transparency, consistency, correctness, correctability. These notions concerns the way data are collected and produced, independently of the distributive outcomes that result.

This of course does not exhaust all logical possibilities. A principle may define a distribution as just based on both information about the process and the end-state of the process (or of that similar processes have in general, or of counterfactual processes). For example, Rawls claims that the distribution of goods resulting from voluntary market exchanges are always just, provided that the market operates within political institutions that have certain specific features. These institutions – e.g. redistributive taxation – are defined by means of the end-states that they reach (e.g. higher expectations for the worst-off individuals than any alternative taxation scheme) (Rawls 1999). Indirect utilitarianism defines justice as the result of rules that maximize utility in the aggregate and over the long term. It does not require that every single implementation of a rule maximizes general utility. For example, freedom of expression has been defended in this way (Mill 1987). We shall refer to these principles as 'procedural-indirect'. In the field of data justice, the 'procedural indirect' seems analogous to the 'distributive justice' approach described by Heeks and Renken (2018), which

defines data justice as the result of data transactions protected by rights (of access to data or ownership of the data), which are justified by reference to their distributive consequences.

Moreover, principles of justice can differ relative to their scope. For example, justice may concern the distribution of goods in the nation-state (domestic justice, e.g. concerning the income distribution in Belgium), within members of voluntary association (local justice, e.g. justice between participants to an online platform), or some broader set, for example the world (global justice) (Rawls 1999).

Finally, principles of justice can differ in their level of idealization relative to the actual world. Rectificatory justice, for example, presupposes an antecedent injustice - it addresses the actual world which contains injustice. Ideal theory justice, as defined by Rawls, concerns the principles of justice that ought to regulate relations between agents, all of whom are assumed to comply with them (Rawls 1999; Van Parijs 2007).

4.2. A general framework of procedural justice for online platforms

In what follows, we propose to assess the justice of nudges as an instance of local justice, namely justice concerning social cooperation of all individuals involved in the operation of the online platform. We shall assume domestic justice: that is, we assume that the fundamental rights of citizens are respected and that society has adequate institutions to ensure that whatever norms of justice apply at the domestic justice level are satisfied. Notice that we make this assumption for the sake of simplicity. A platform that operates in an unjust context, e.g. in a country in which fundamental rights are not respected, may have to operate by slightly different principles. This paper however pursues a more limited goal. The simplifying assumption requires us to focus on the problems of justice that are located entirely *within the platform,* because of the platforms. Also, we abstract here from the question whether online platforms undermine or even functionally replace (Loi and Dehaye 2016) social institutions necessary to achieve justice in society, even though we believe this question to be extremely important. Third, we adopt a procedural-indirect view of justice. That is, we regard any informational and economic outcome (e.g. distribution of data, attention, resources) as just that is achieved through exchanges and communication which respects the rules and nudges of well-ordered platforms. We characterize the rules and nudges of well-ordered platforms by considering the goals such rules and nudges achieve. Fourth, we select as the currency of justice two types of platform primary goods:

- *intrinsic platform goods*: these are identified with the official goals for the sake of which a platform claims to exist. For example, these goals are “to build community and bring the world closer together” in the case of Facebook (2019), or “to organize the world’s information and make it universally accessible and useful” for Google (2019). In ideal circumstances, these goals are also meaningfully related to the goals platform users try to achieve through them.
- *platform powers and responsibilities*: these are capabilities to control and influence outcomes, resulting from the rules of the platforms and its software architecture.

We ignore the goods of income and wealth, even platforms affect their distribution, because we assume that in a just society there are institutions that ensure achieving a just distribution of these goods anyway (e.g. thanks to taxation and redistribution, welfare, equality of opportunity in education, etc).

We shall assess the justice of nudges by appealing to three principles of justice that are individually necessary and jointly sufficient conditions for platforms to be just. These three principles are:

First principle of platform justice (harm principle)*:*
Choice architectures in online platforms are just
i. *only if* they do not significantly increase risk of significant harm to which any user is exposed and
ii. *if* (i.e. *not* only if) they reduce the (aggregative) risk of significant harm to which all users are exposed.

Second principle of platform justice (non-discrimination principle)
Unequal prospects of obtaining intrinsic platform goods, which result from the expression of preferences of other users in the platform are just only if users who meet the *relevant* preferences of other users to a similar degree have similar prospects of earning intrinsic platform goods, irrespective of their unequal *irrelevant* traits (such as, depending on the context, sex, gender, race, ethnicity, religion, etc).

Third principle of platform justice (benefit principle)*:*
inequalities in power and responsibility between different platform roles are just only if a less unequal distribution would be less beneficial for the role that has the least power.

The three principles are hierarchically ordered, that is, the n^{th} principle must be satisfied before the $n^{th}+1$ is satisfied. A platform that satisfies them is well-ordered. In a well-ordered platform, if inequalities emerge between platforms users by virtue of the preferences expressed by other users (e.g. some users sell more stuff, have more followers, have larger social networks), these inequalities are just. Assuming that is the case, platforms ought to respect the outcomes of the actual expression of the preferences of platform users, enabled by the platform software architecture, even if the outcomes are unequal. Justice is preserved even if preferences are nudged, *provided* all principles are satisfied.

The first principle of platform justice uses risk of significant harm to refer to significant increases in the probability of an event that risks undermining the security, subsistence, health, or dignity of the user. Notice that nudges which other principles justify are prohibited if they significantly increase the risk of significant harm for even a single user by virtue of (i). Conversely, nudges that reduce the risk for one or more individuals (ideally, all individuals), without, at the same time, increasing the risk of any other individuals, are also justified as they *jointly* satisfy (i) and (ii).

An example of an event that undermines the security of a user, is access to confidential information (e.g. geographical location) about a user who wishes to keep this information confidential, by a person or state entity willing to exercise violence against that user (e.g. for political reasons). An example of a subsistence risk is malicious access to a person's financial resources, which may be diverted away and cause significant economic hardship. An example of health risk is severe addiction, leading to a diagnosed pathology. Two examples of dignity risk are access to sensitive information, which is then used to compromise a person's reputation, and online identity theft. Minor threats of harm (harm defined as 'having less of anything that is a good') are not pertinent to this principle. The principle has to be limited to important harms, otherwise the satisfaction of this principle becomes too demanding, and incompatible with most nudges, which often make it easier to

obtain a good (e.g. attention) for someone while making it harder to obtain the same, or another, good for someone else. By contrast, the distribution of other benefits is governed by the other two principles. With reference to Taylor's three pillars' scheme, the first principle concerns the first pillar (privacy) and extends it, for instance, so that concerns of cybersecurity are also considered.

The second and third principles are both inspired by distinct parts of Rawls's (1999) second principle of justice. The second principle is a generalization of the second part of Rawls's second principle, the fair equality of opportunity principle (FEO). FEO deals with prospects of obtaining social positions in society, because it is principle for the fundamental social institutions of a society, and one of the most important questions of social justice concerns the procedures through which individuals acquire their social positions. By contrast, the second principle of *platform* justice focuses on *prospects of obtaining intrinsic platform goods*. The relevant equality requirement concerns not equality among the equally talented and motivated, as in Rawls (1999), but, instead, equality among those who are equal with respect to 'whatever fits the *relevant* preferences of users in the platform'. *Relevant* is the important word here. The principle does not require absolute equality of prospects but is compatible with unequal prospects for users who are unequal in the relevant respects. Not all inequalities that are addressed by the preferences of other users should be considered relevant. Racist preferences for example should not be considered relevant in a platform that distributes job opportunities, or rental places. What is or not relevant for each platform our principles do not say: it is a normative choice to be made by using ethical resources outside those specified by our theory. Clearly it depends from what the platforms are for. With reference to Taylor's (2017) three pillars, the second principle belongs to the third pillar, non-discrimination.

The third principle focuses on the different platform roles involved in platform interactions. Platform roles are functional social roles that enable the platform to exist and, conversely, only exist because there are platforms. For example, a platform Chief Operating Officer, the user (often, the different kinds of users), the employee (normally, different kinds of employees, such as product designer), the owner, the shareholders, etc. Typically, each role corresponds to contractually defined (legal) rights and responsibilities. Different rights and responsibilities imply that different platform roles enjoy different amounts of powers. Not all powers are, however, stated in legal terms. Some powers emerge not as a result of contract but of a platform's software *architecture* (Lessig 2006). For example, the chief technology officer may exert indirect power over the platform's users by virtue of her power to design non-educative nudges, influencing the choices of users. As argued in part 3, non-educative choice architectures exert influence on the nudged person, which does not take engage sufficiently the nudged person's autonomy. We consider the 'power to nudge' as a *power* that is distributed unequally between those who design choice architectures and those who act through them. This unequally distributed power is also the chief inequality which concerns us in this paper.[4] This is the subject matter regulated by the third principle of platform justice. It specifies what must obtain for this unequally distributed power to be justified. The second principle of platform justice does not require that equal power between platform designers and users – that is, no non-educative nudges. It justifies an unequal distribution which can be functionally justified, when the functional justification fulfills a principle of reciprocity inspired

[4] By contrast, educative nudges engage with the subject's reflective capacities. The user retains full control and, for that reason, educative nudges do not count as unequally distributed powers that require a justification.

by Rawls's Difference Principle. Reciprocity is justified by inequality when it is necessary to benefit individuals in the least powerful social role (Rawls 1999) – which relative to nudges is that of the platform user. The concept of *benefit* that is the currency of this principle refers to *intrinsic platform goods*.[5] The distribution of wealth and income is not a concern of platform justice, but a concern of domestic justice, which we for simplicity assume to be achieved. With reference to Taylor's scheme, this belongs to the second pillar, engagement with technology, because nudges that make it easier for users to obtain the good of a platform contribute to Taylor's engagement, and sharing the data benefits in a more inclusive manner. E.g. data contribute to make technology accessible for free, thus to the poorest citizens, as claimed by Google. But notice that such justification is valid only if it does not violate the other two, hierarchically higher, principles.

4.3. Justice and digital nudges

The three principles of platform justice are relevant to evaluating choice architectures. In what follows, we provide some illustrations of how the three principles may be applied to existing platforms to identify injustices within them and recommend improvements.

Dimension 1: reviewing and choosing privacy and data sharing options

The general principle that seems to find immediate and relevant application here is the third principle of platform justice. That implies that nudges for making private facts accessible and for sharing more data are just only if necessary to improve the users' expectations of intrinsic platform goods. One implication would be that, on a social network like Facebook, nudges for sharing data can be justified if they connect individuals and communities. By contrast, nudges to provide data, that are not necessary to enhance connections between people, but just because the platform benefits from them, are not justified.

Of course one big problem of this criterion is that the intrinsic goods officially enabled by platforms are to some extent fuzzy and, in addition to that, they may change over time. Consider Facebook's collections of data to favor sexual encounters. Are sexual encounters included in Facebook's intrinsic good of community and a more connected world? There might be reasonable disagreement concerning this. At any rate, there has to be coherence between the good a company officially pursues and the use of the data it collects through nudges. Nudges that are not necessary to promote those goods are not justified. Moreover, even nudges that are necessary to promote those goods are not justified, if they violate the second and first principle, i.e. if they are discriminatory or pose significant threats of significant harm.

Dimension 2: selecting the beneficiaries of platform interactions

Here we discuss nudges that influence the selection of beneficiaries of platform interactions. Let us consider, as an example, a platform for renting own apartment spaces (e.g. beds, rooms or whole houses) between the platform's own users. The second principle requires that the choice architecture is designed in such a way that users that meet the relevant preferences of other users to a similar degree will have similar prospects of earning intrinsic platform goods, irrespective of unequal irrelevant features. In the context in question,

[5] The principle is coherent with Sunstein's idea that system 1 nudges respect autonomy when they facilitate the satisfaction of preferences of system 2. But it is also narrower, for it does not refer to a user's welfare, or preferences, *in general* but to the goods and preferences relevant for particular types of platform interactions. These are the goods officially pursued or enabled by platforms, those that it is the mission of the platform to promote or generate.

the intrinsic platform goods are apartment spaces, for people seeking space, and economic opportunities, for people with a place to rent. Relevant preferences should be considered, in this context, preferences related to the quality of the accommodation, e.g. its location, cleanness, available facilities, etc. A preference for the race of the person renting the space should not be considered relevant. In the competition for offering space, the principle requires the choice architecture to be designed in such a way that users offering space who meet the relevant preferences of users seeking space should have the same economic opportunities irrespective of, for example, their race. Which features should be considered relevant and irrelevant varies depending on the context. In a platform for sharing apartment space to strangers, it may be argued that discrimination based on sex or gender should be allowed. On the room seeker side, people with similar relevant traits (e.g. record of kindness and care for the property) should have the same prospects of obtaining the intrinsic goods (apartment spaces) irrespective of their non-relevant features (e.g. race). Choice architectures that violate this conditions are unjust; choice architectures that repair or prevent this condition are just (provided they do not violate the first principle).

Dimension 3: Cybersecurity

For cybersecurity, the first principle of platform justice is obviously the most relevant one, coupled with the third. The first principle clause (i) limits the nudges that the second and third principle otherwise justify. The third principle justifies all nudges contributing to improving the users' expectations of intrinsic platform goods. These may be nudges that make it easier, but also more insecure, for people to log in in their accounts. The third principle may justify these insecure nudges in a social networking platform, because, for example, the easier for people to join a platform, the more people actually join it, and the more people join it, the higher the likelihood of connecting people. The first principle prohibits insecure design choices, e.g. because they pose a threat to dignity, as they raise the risk of identity theft. Losing control over one's online identity online can be considered a threat to dignity, so a choice architecture that increases its possibility violates the first principle. Conversely, all nudges that are required for cybersecurity are justified, even if they are not necessary to augment expectations of intrinsic platform goods. Similarly, nudges that prevent individuals to develop addictions are justified, even if they reduce the goods otherwise made available on the platforms.

Dimension 4. Horizontal expansion of services

Let us now consider nudges to install services delivered through specific providers. The first principle of platform justice entails that nudges are just if they reduce risk of significant harm to the users and only if they do not increase the risk of significant harm of users. Significant harm refers to threats to security, subsistence, health, or dignity. When this and the non-discrimination principle is satisfied, the third principle justifies nudges to horizontally expand services that are necessary to augment the users' expectations of the intrinsic platform goods. So, for example, a nudge to use Apple pay on Apple products is just if it makes financial transactions more secure, or if it is necessary to improve the usability of the services (while not exposing the user to higher cyber threats). In evaluating the satisfaction of the first principle (i.e. whether the nudge makes the user more secure), the relevant baseline are the payment services of other competitors, that users could use. If the use of Apple pay does not guarantee higher security, the nudge violates the second principle, because it discriminates based on irrelevant features. So in the latter case, users should not be nudged.

Dimension 5. Political power of the platform

Prima-facie the dimension of political power - nudging users to address their political representatives - may appear always problematic from the point of view of justice. Take the case of Uber. Is the use of nudges to address political representatives who want to regulate

Uber justified? One can appeal to the third principle of justice to explain what has to be the case for this to be true: it must be the case that this is *necessary* to improve the users' expectations of Uber's intrinsic platform good, which is, as they define it, to bring transportation for everyone, everywhere. Uber's managers could argue that the regulation of Uber is inimical to this goal, so the nudge to activate a political campaign against it, serves the goal. But suppose that Uber would nudge its drivers to accept rides from people joining the political campaign. That would violate the second principle, that has priority relative to the third. Or consider the Facebook case. According to the third principle of platform justice, the 'get out the vote' and similar initiatives by Facebook are only justifiable if political participation, or engagement, is an intrinsic platform good for Facebook, which seems plausible given that political participation is a significant aspect of what brings people together. However, some Facebook initiatives may violate the Second Principle of Platform Justice. The relevant contrast here is between an initiative like the 'get out the vote' and one like adding a rainbow reaction to signal support to LGBT causes. This could violate the principle that individuals with the same capacity to meet the preferences of other users have similar prospects of obtaining intrinsic platform goods. If would be violated if, for example, supporters of communist causes, or conservative causes, or animal rights causes, or environmental causes, have lesser chances of building political communities because they lack a similar nudge. Similarly, it is not necessarily problematic for a political participation platform to nudge the active engagement of its users. The nudge is as such a form of unequal power to the platform designers, but this is justified if it fulfils all principles, including being necessary to benefit user in terms of the intrinsic platform good, namely political engagement. It is however problematic to build nudges that confer advantage 'by design' to some users, compared to others, or to some political decisions, above others. For example, in the Rousseau platform - the online platform for the political deliberation and voting of the Italian *Movimento Cinque Stelle* - nudges that influence users to vote are justified, but nudges that influence users to vote *in a certain way* are not.

Dimension 6 policing of social norms

As mentioned in the analysis of this dimension, above, here we are dealing with nudges that affect the degree of civility in the online communication environment of users. Nudging for more formal politeness and courtesy is always justified because it enhances the fruition of any intrinsic good: in any platform that involves communication between users with each other or with the platform employees, fruitions of the goods the platform provides is facilitated by such norms. What can be morally problematic, and is more interesting to assess morally, is the nudging of views that may be deemed offensive, even when they are expressed politely, or content, including visual content, that is regarded to be obscene. The vagueness of social norms concerning what counts as offensive or hate speech, or visual obscenity entails that here the platform deals with fuzzy boundaries and threshold have to be set arbitrarily. The problematic nudges are those affecting the degree to which borderline content is polished. What makes such nudges problematic is the existence of a trade-off. We define norms for polishing content 'stricter' if and only if they are overall more likely to elicit social sanctions compared to other social norms. Such social sanctions can be expressed through elements of the choice architectures – e.g. buttons for reporting obscene content. Different web-designs may make it easier for users to flag content deemed improper. Online design elements may influence the users' sensitivity to borderline content, e.g. how likely they are to flag something as inappropriate, and the range of possible reactions to them. This kind of choice architecture involves a trade-off between distinct human goods that may be promoted, to a different degree, by stricter or less strict norms. Stricter norms imply a higher proportion of 'false positives' - contents that are flagged as inappropriate and removed from platforms, in spite of

their potential contribution to some value. Less strict norms favour freedom of speech in some way, even if when views deemed offensive and obscenity may also have a chilling effect on users. The perceptual salience design elements for reporting improper content (e.g. flagging buttons) may be designed to influence communication one way or another.

Can such (non-educative) nudges be justified? The first principle of justice justifies these nudges when they are needed to preserve the dignity of participants. But the problem is that there are instances of speech that gets flagged erroneously, or maliciously, which are not truly offensive. The perceptual prominence of flagging symbols may support a social environment that instead of expressing tolerance to borderline content and unpopular opinions, uses all means at its disposal in order to silence undesired ideas. The third principle of justice requires that nudges are necessary to benefit users as measured by the intrinsic good of the platform. So, the amount of nudging that is appropriate, and its direction (whether to promote stricter polishing or more liberal speech) always depends on the purpose of the platform. In a platform for political deliberation, for example, nudges that lead to the policing of opinions will only be justified if they are necessary to avoid chilling the participation of most users. But nudges facilitating the reporting of content as inappropriate, leading to an emotionally safer, more protected environment, could be justified for a platform that serves a different goal.

Conclusions

We have discussed six classes of nudged actions by online platform. Following Thaler's and Sunstein's distinction between educative and non-educative nudges we have argued that educative nudges are always morally permissible, while non-educative nudges are *pro-tanto* morally wrong, but not wrong all things considered. We argue that a platform involving non-educative nudges is well-ordered when three principles of platform justice are satisfied, even if non-educative nudges are *pro-tanto* wrong because they do not engage the users' capacities for autonomous choice. We illustrate the plausibility of the three principles by applying them to the six classes of nudges distinguished above.

The moral analysis of nudges provides an indirect-procedural account data justice, alternative to accounts that focus on merely the inequality created by data in some broad currency of justice (e.g. well-being, capabilities, wealth, etc) or focus on information norms as disciplinary tools or expression of social privilege (Johnson 2014) or a historical-proceduralist justice that focuses on fair data handling (Mann 2016). It bears similarities to a distributive justice approach focused on rights (Mann 2016) justified by their effects.

The first kind of account faces the problem that it is unreasonable to expect online platforms (in general) to generate all things considered outcome equality (or maximization, etc…) in overall well-being, or wealth, or income, or power, etc. Most platforms, after all, only affect a limited aspect of people's lives. (Theoretically, the exceptions are 'dominant' platforms, those that become so pervasive that social interactions through them become unavoidable, see Loi and Dehaye 2016.) For most platforms it is extremely difficult to assess how they influence the distribution of overall well-being and other important social goods in society. The second type of account, focused on disciplinary power or social privilege cannot tackle systems that are both highly inclusive and not based on coercion, as many online platforms are. Our proceduralist approach applies to these, and characterizes the distribution and use of data as *unjust* if it results from online interactions nudged by a choice architecture which violates the three principles of platform justice in hierarchical order.

A limitation of this approach is that it rests on the idealization that the basic structure of society (Rawls 1999) is just. That is, it assumes that basic rights and liberties are respected and that institutions are in place to achieve background justice, i.e. a just distribution of the most important social goods and opportunities. In actual fact, there is no *a-priori* guarantee that well-ordered platforms will not undermine the functioning of social institutions that are necessary for background justice. But that is the topic for another article.

References

Bicchieri C., Dimant E. 2019. Nudging with Care: The Risks and Benefits of Social Information. SSRN, April 19. http://dx.doi.org/10.2139/ssrn.3319088.

Blumenthal-Barby, J.S. 2016. Biases and Heuristics in Decision Making and Their Impact on Autonomy. The American Journal of Bioethics, 16, 5: 5-15.

Bohn, Dieter. 2015. “Uber Trolls New York City Mayor in Its App.” The Verge. July 16, 2015. https://www.theverge.com/2015/7/16/8981015/uber-trolls-new-york-city-mayor-de-blasio-app.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. “A 61-Million-Person Experiment in Social Influence and Political Mobilization.” Nature489 (7415): 295–98. https://doi.org/10.1038/nature11421.

Bovens, L. 2009. The Ethics of Nudge, in Preference Change. Approaches from Philosophy, Economics and Psychology. eds. T. Grüne-Yanoff and S.O. Hansson 206-19. Dordrecht: Springer.

Briggs, P., D. Jeske, and L. Coventry. 2017. “Behavior Change Interventions for Cybersecurity.” In Behavior Change Research and Theory, edited by Linda Little, Elizabeth Sillence, and Adam Joinson, 115–36. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-802690-8.00004-9.

Brink, D. 2013. Mill’s Progressive Principles. Oxford: Clarendon Press.

Chakraborty, Abhijnan, Nuno Mota, Asia J. Biega, Krishna P. Gummadi, and Hoda Heidari. 2019. “On the Impact of Choice Architectures on Inequality in Online Donation Platforms.” In . ACM. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3027017.

Coons C., Weber M., eds. 2014. Manipulation: Theory and Practice. Cambridge: Cambridge University Press.

Coventry, Lynne, Pam Briggs, Debora Jeske, and Aad van Moorsel. 2014. “SCENE: A Structured Means for Creating and Evaluating Behavioral Nudges in a Cyber Security Environment.” In Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience, edited by Aaron Marcus, 229–39. Lecture Notes in Computer Science. Springer International Publishing.

Dolan, P., M. Hallsworth, D. Halpern, D. King, R. Metcalfe, and I. Vlaev. 2012. “Influencing Behaviour: The Mindspace Way.” Journal of Economic Psychology33 (1): 264–77. https://doi.org/10.1016/j.joep.2011.10.009.

Donaldson, Thomas, and Thomas W. Dunfee. 1999. Ties That Bind: A Social Contracts Approach to Business Ethics. Harvard Business School Press.

Facebook. 2019. “Facebook - Resources.” Accessed May 8, 2019. Available from https://investor.fb.com/resources/default.aspx.

Goodwin, T. 2012. Why We Should Reject ‘Nudge’. Politics, 32, 2: 85-92.

Google. 2019 “About” Access May 8, 2019. Available from: https://about.google/intl/en/

Gorin M. et al. 2017. Justifying Clinical Nudges. Hastings Center Report, 47, 2: 32-8.

Griffin, J. 1986. Well-Being. Its Meaning, Measurement, and Moral Importance. Oxford: Clarendon Press.

Hansen P.G., Jespersen A.M. 2013. Nudge and the Manipulation of Choice. A Framework for the Responsible Use of the Nudge Approach to Behavior Change in Public Policy. European Journal of Risk Regulation, 4, 1: 3-28.

Hartzog, Woodrow. 2018. Privacy’s Blueprint: The Battle to Control the Design of New Technologies. Cambridge MA: Harvard University Press.

Hausman D.M., Welch B. 2010. To Nudge or Not to Nudge. The Journal of Political Philosophy, 18, 1: 123-36.

Hertwig R., Grüne-Yanoff T. 2017. Nudging and Boosting: Steering or Empowering Good Decisions. Perspectives on Psychological Science, 12, 6: 973-86.

Holm S. 2017. Authenticity, Best Interest, and Clinical Nudging. Hastings Center Report, 47, 2: 38-40.

Lessig, L. 2006. Code, 2.0 New York: Basic Books.

Levin, Sam, and Julia Carrie Wong Luke Harding in London. 2016. “Facebook Backs down from ‘napalm Girl’ Censorship and Reinstates Photo.” The Guardian, September 9, 2016, sec. Technology. https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

Loi, Michele, and Paul-Olivier Dehaye. 2017. “If Data Is the New Oil, When Is The Extraction of Value From Data Unjust.” Philosophy and Public Issues 7 (2): 138–78.

Mann, Laura. 2018. “Left to Other Peoples’ Devices? A Political Economy Perspective on the Big Data Revolution in Development.” Development and Change 49 (1): 3–36.

Mill, John Stuart. 1987. “Utilitarianism.” In Utilitarianism and Other Essays, edited by Alan Ryan, 272–338. Harmondsworth, Middlesex, England; New York, N.Y., U.S.A.: Penguin Books.

Nozick, Robert. 1974. Anarchy, State, and Utopia. New York: Basic Books.

NTB. n.d. “Norsk Forfatter Midlertidig Utestengt Fra Facebook Etter å Ha Postet Bilde Fra Vietnamkrigen.” Aftenposten. Accessed October 26, 2016. http://www.aftenposten.no/article/ap-603854b.html.

Nys, T.R.V., Engelen, B. 2017. Judging Nudging: Answering the Manipulation Objection, 65, 1: 199–214.

O’Neill, Onora. 2000. Bounds of Justice. Cambridge University Press.

Rawls, John. 1999. A Theory of Justice. 2nd ed. Cambridge, MA: Harvard University Press.

Redmiles, Elissa M., Michelle L. Mazurek, and John P. Dickerson. 2018. “Dancing Pigs or Externalities?: Measuring the Rationality of Security Decisions.” In Proceedings of the 2018 ACM Conference on Economics and Computation, 215–232. ACM.

Reijula S. et al. 2018. Nudge, Boost, or Design? Limitations of Behaviourally Informed Policy under Social Interaction. Journal of Behavioral Economics for Policy, 2, 1: 99-105.

Schmidt, A.T. 2017. The Power to Nudge. American Political Science Review, 111, 2: 404-17

Schwab, A.P. 2006. Formal and Effective Autonomy in Healthcare. Journal of Medical Ethics, 32: 575-9.

Sjefredaktør, Espen Egil Hansen. n.d. "Dear Mark Zuckerberg. I Shall Not Comply with Your Requirement to Remove This Picture." Aftenposten. Accessed October 26, 2016. http://www.aftenposten.no/article/ap-604156b.html.

Sunstein C.R. 2014. Nudging: A Very Short Guide. Journal of Consumer Policy, 37, 4: 583-8.

Sunstein C.R. 2016. The Ethics of Influence. Government in the Age of Behavioral Science. Cambridge: Cambridge University Press.

Taylor, Linnet. 2017. "What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally." Big Data & Society4 (2): 2053951717736335. https://doi.org/10.1177/2053951717736335.

Temperton, James. n.d. "Facebook Makes U-Turn on Decision to Censor an Iconic Vietnam War Photo." WIRED UK. Accessed October 26, 2016. http://www.wired.co.uk/article/facebook-terror-of-war-vietnam-napalm-girl-image-censored.

Thaler R.H., Sunstein C.R. 2011. Nudge. Improving Decisions about Health, Wealth, and Happiness. New Haven: Yale University Press.

Van Parijs, Philippe. 2007. "International Distributive Justice." In A Companion to Contemporary Political Philosophy, edited by Robert E. Goodin, Philip Pettit, and Thomas Pogge, 2:638–52. Oxford: Blackwell. http://www.ucl.be/cps/ucl/doc/etes/documents/InternationalDistr.Justice.pdf.

Weinmann, Markus, Christoph Schneider, and Jan vom Brocke. 2016. "Digital Nudging." Business & Information Systems Engineering58 (6): 433–36. https://doi.org/10.1007/s12599-016-0453-1.

Wong, Julia Carrie. 2016. "Mark Zuckerberg Accused of Abusing Power after Facebook Deletes 'napalm Girl' Post."The Guardian, September 9, 2016, sec. Technology. https://www.theguardian.com/technology/2016/sep/08/facebook-mark-zuckerberg-napalm-girl-photo-vietnam-war.