sustainability

MDPI

Article

The Ethics of AI-Powered Climate Nudging—How Much AI Should We Use to Save the Planet?

Marius Bartmann

German Reference Centre for Ethics in the Life Sciences (DRZE), University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany; bartmann@uni-bonn.de

Abstract: The number of areas in which artificial intelligence (AI) technology is being employed increases continually, and climate change is no exception. There are already growing efforts to encourage people to engage more actively in sustainable environmental behavior, so-called "green nudging". Nudging in general is a widespread policymaking tool designed to influence people's behavior while preserving their freedom of choice. Given the enormous challenges humanity is facing in fighting climate change, the question naturally arises: Why not combine the power of AI and the effectiveness of nudging to get people to behave in more climate-friendly ways? However, nudging has been highly controversial from the very beginning because critics fear it undermines autonomy and democracy. In this article I investigate the ethics of AI-powered climate nudging and address the question whether implementing corresponding policies may represent hidden and unacceptable costs of AI in the form of a substantive damage to autonomy and democracy. I will argue that, although there are perfectly legitimate concerns and objections against certain forms of nudging, AI-powered climate nudging can be ethically permissible under certain conditions, namely if the nudging practice takes the form of what I will call "self-governance".

Keywords: sustainability; climate change; artificial intelligence; nudging; digital nudging; libertarian paternalism; autonomy; intergenerational justice

check for updates

Citation: Bartmann, M. The Ethics of AI-Powered Climate Nudging—How Much AI Should We Use to Save the Planet? *Sustainability* **2022**, *14*, 5153. https://doi.org/10.3390/su14095153

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 28 February 2022
Accepted: 22 April 2022
Published: 24 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The number of areas in which artificial intelligence (AI) technology is being employed increases continually, and climate change is no exception [1]. Several challenges posed by AI have been discussed extensively over the past years—privacy, bias, opacity, to name but a few examples [2]. More recently, the environmental impact of AI itself has also come into focus, for example, the high energy consumption needed to train and run algorithms [3,4]. This research contributes to the insight that we must be wary not to indulge in a questionable form of "technological solutionism" [4] (p. 71) that views AI as a panacea for all kinds of problems. Thus, we must carefully balance the costs and benefits of AI's employment, with particular scrutiny of the ethical challenges involved [5,6].

There are already growing efforts to encourage people to engage more actively in sustainable environmental behavior, so-called "green nudging", such as certifying consumer products with eco-labels and providing households with peer comparisons to improve energy conservation [7]. Nudging in general is a widespread policymaking tool designed to influence people's behavior while preserving their freedom of choice [8–10]. With people spending more time in the digital sphere and making more decisions online, nudging has also become widespread in digital environments [11–13]. Big data and AI are being used in the service of "Big Nudging" to steer people's choices [14,15]. Given the enormous challenges humanity is facing in fighting the severe, harmful and possibly irreversible effects of climate change [16], the question naturally arises: Why not combine the power of AI and the effectiveness of nudging to get people to behave in more climate-friendly ways? Climate nudging powered by AI may thus suggest itself as a suitable strategy to

change people's behavior so that their decisions contribute to a cleaner, safer, and more sustainable planet.

However, nudging has been highly controversial from the very beginning [17]. Advocates praise it as an effective means for making people's lives better. Critics object that nudging compromises people's autonomy by interfering with their capacity to make their own choices. Employing AI to nudge people is no less controversial, even if it is being done with the well-intentioned effort to prevent further harmful climate change. Critics reject AI-powered nudging as large-scale paternalism, which not only disrespects people's autonomy but may also lead to authoritarian societies in which a digital "Green Leviathan" [15] or "wise king" [14] manipulates our lives behind our backs.

In this article I investigate the ethics of AI-powered climate nudging and address the question whether implementing corresponding policies may represent hidden and unacceptable costs of AI in the form of a substantive damage to autonomy and democracy. I will argue that, although there are perfectly legitimate concerns and objections against certain forms of nudging, AI-powered climate nudging can be ethically permissible under certain conditions. To this end, I first elaborate on nudging in general, its background, and rationale (Section 2). In a second step, I review the main argument for nudging as well as the issue of autonomy as the main ethical concern critics have raised (Section 3). Third, I briefly present relevant facts about climate change and the specific ethical challenges they raise (Section 4). Finally, I consider AI-powered climate nudging and discuss whether the main ethical concerns revolving around autonomy also apply in this context. I argue that AI-powered climate nudging can be ethically permissible if the nudging practice takes the form of what I will call "self-governance" (Section 5).

2. Libertarian Paternalism and Nudging

In their highly influential 2008 book *Nudge*, Richard Thaler and Cass Sunstein develop a policymaking approach they call *libertarian paternalism*:

> We strive to design policies that maintain or increase freedom of choice. When we use the term *libertarian* to modify the word *paternalism*, we simply mean liberty-preserving. And when we say liberty-preserving, we really mean it. Libertarian paternalists want to make it easy for people to go their own way; they do not want to burden those who want to exercise their freedom. The paternalistic aspect lies in the claim that it is legitimate for choice architects to try to influence people's behavior in order to make their lives longer, healthier, and better. [8] (p. 5)

Early on, commentators have noted that the definition of paternalism given by Thaler and Sunstein deviates significantly from standard accounts [18] (pp. 126–130). Compare their definition with a standard definition of paternalism:

> Paternalism is the interference of a state or an individual with another person, *against their will*, and defended or motivated by a claim that the person interfered with will be better off or protected from harm. [19] (emphasis added)

In both definitions, increasing an agent's individual welfare represents the primary aim of the interference and serves at the same time as a justification for it. The difference consists in the means employed to achieve this aim. According to the standard account of paternalism, interferences to increase individual welfare infringes on the liberty or autonomy of the targeted person in some way or other. More often than not, the infringement consists in altering or restricting the space of options among which people can choose, for example, banning smoking in public buildings or prescribing motorcyclists to wear helmets [19]. Libertarian paternalism, on the other hand, purports to increase individual welfare while respecting people's liberty and autonomy by preserving freedom of choice. Libertarian paternalism leaves the space of options intact, that is, the interference consists not in *what* options are made available but rather in *how* the options are presented. A standard example is a cafeteria where salads, fruits, vegetables, and other healthy food items are deliberately displayed at eye level so that visitors are more likely to choose what is

better for them [8] (pp. 1–4). This is precisely what Thaler and Sunstein call a "nudge", the notion at the heart of libertarian paternalism. It is defined in the following way:

> A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not. [8] (p. 6)

The term "choice architecture" refers to the space of options within which people make choices, and nudges are intentional modifications of this space to influence people's choices for their own benefit in a way that does not interfere with their freedom of choice. The number of options is supposed to remain constant, only the presentation of options is changed. A key feature of standard paternalism frequently raising ethical concerns—interfering with people's liberty or autonomy in some form—is thus declared absent from libertarian paternalism.

Before elaborating on nudging and choice architecture in more detail, I want to emphasize right from the outset a rather general point often highlighted by the authors themselves. Sunstein calls this general point the "trap of abstraction" [10] (p. 424). The point is plain and simple but important: nudges can assume a vast and diverse variety of forms; they involve different ends, different means, and different justifications. Hence, a proper ethical assessment of nudges should always consider their characteristic features and the specific circumstances of their implementation.

To get an impression of the heterogeneity of nudges, consider the following examples given by Sunstein himself [10] (p. 424). A nudge can consist in merely providing consumers with information or warnings on products (a GPS device, nutritional information on food items, or graphic images on cigarette packages), in reminders for unpaid bills, in rearranging items in the supermarket to increase their salience, or in changing the default from opt-in to opt-out in the context of enrollment in savings or retirement plans. Paradigmatically, nudges like this aim at people's individual welfare, they are intended to promote people's own ends and to make their lives better. Unsurprisingly, that is why the issue of paternalism has taken center stage in the nudging debate. Yet many nudges are not paternalistic in that they rather aim at social welfare or at protecting the environment, for example, campaigns and programs to reduce energy consumption and greenhouse gas (GHG) emissions [8] (chapter 12). Since the purpose of nudges benefitting the environment is to correct market failures, Sunstein later on expressly distinguishes between "paternalistic nudges" and "market failure nudges" [10] (pp. 426–427).

The upshot of the general point is this: the ends, means, and justifications constituting the respective nudges can differ substantially, therefore the ethical questions involved and the ethical assessment required to decide them may be just as different. The lesson to be drawn is that rather than trying to arrive at a general verdict as to the ethical permissibility of nudging per se we are better off proceeding in a somewhat piecemeal fashion and discuss relevant ethical principles always with an eye towards specific nudging practices.

Now, what is the rationale for libertarian paternalism and nudging in the first place? Libertarian paternalism is based on a certain picture of human cognition. More specifically, it is rooted in a particular understanding of the mechanisms underlying decision-making processes. Thaler and Sunstein largely draw on behavioral science research, especially on the works of Daniel Kahneman [8] (chapter 1). According to Kahneman, the human mind contains two cognitive systems that process information in significantly different ways. System 1 "operates automatically and quickly", whereas System 2 "allocates attention to the effortful mental activities that demand it, including computations" [20] (pp. 20–21). System 1 is in charge of activities such as sensing a particular mood in a voice, driving on an empty highway, or processing simple sentences; System 2 activities require concentration and focus, such as following the clowns in a circus show, trying to remember a particular sound, or doing the taxes [20] (pp. 21–22). Although both systems are fallible, System 1 is particularly susceptible to what is nowadays known under the umbrella term "heuristics

and biases" [20] (part II). Among the many lapses and blunders of System 1 Thaler and Sunstein point out are, for example, overconfidence, loss aversion, and framing [8] (chapter 1). Take one of the most famous experiments illustrating the framing fallacy. Physicians were given information about a certain medical procedure. Those who read the description "The one-month survival rate is 90%" were much more likely to decide in favor of the procedure than those physicians who read the description "There is 10% mortality in the first month", even though both descriptions are logically equivalent [20] (p. 367). Examples like these and the fallacies associated with System 1 abound.

Thaler and Sunstein conclude from the ubiquitous shortcomings of System 1 that the *homo economicus* promulgated by many economists is an illusion. Choices and decisions are simply not always the outcome of fully rational, fully informed, and strong-willed individuals who always act in accordance with their best interests:

> The false assumption is that almost all people, almost all of the time, make choices that are in their best interest or at the very least are better than the choices that would be made by someone else. We claim that this assumption is false–indeed obviously false. [8] (p. 8)

For example, they point to the scientifically corroborated link between obesity and the increased risk of several medical conditions on the one hand, and the high obesity rate in the U.S. on the other to cast doubt on the idea that all U.S. citizens keep to an ideal diet [8] (p. 7). So why not take advantage of the weaknesses of System 1, so the reasoning goes, and harness the power of nudges to get people to choose what is better for them? The food items in the cafeteria must be arranged somehow—why not make a virtue of necessity and design the choice architecture in a way that helps people achieve their ends (a healthy lifestyle) and still preserves their freedom (they can choose a less healthy alternative if they want)?

3. Nudgers and Their Critics

Roughly, there are three main arguments for nudging. The first argument simply builds on the effectiveness of nudging (this is largely uncontroversial, and I will not dispute it here); the second argument maintains that nudging preserves freedom of choice and autonomy (this is very controversial, and I will discuss it in the context of climate nudging in Section 5) [8] (p. 252). Here, I want to review briefly the third argument because it is presented by Sunstein and Thaler as one of their central arguments and also sheds light on a core notion of libertarian paternalism: choice architecture.

The argument is that choice architecture is inevitable, therefore nudging is permissible ([8] (p. 237)), ([9] (p. 14)), ([10] (pp. 415, 420–422)). Just as homo economicus was an illusion, so was the neutrality of the space of options within which people make choices. Indeed, this claim seems hard to challenge. The food items in the cafeteria must be arranged somehow, and the information about mortality and survival rates of medical procedures must be worded in some way. Choice architecture cannot but influence people in one way or other regardless of whether it is the result of deliberate design or coincidence. This also applied to government regulation. In many cases, public officials cannot avoid acting as choice architects in policymaking. This was particularly true when it came to defaults, for example in organ transplantation. *Some system* has to be put in place (for example, opt-in or opt-out), and whichever system is chosen will have consequences for people. Since choice architecture is thus inevitable, nudging is permissible.

It's unclear to me whether the conclusion follows from the premise. Even if the premise is true, and people are influenced by choice architecture regardless of whether it is the result of deliberate design or coincidence, does it really follow that nudging is ethically permissible per se? All that seems to follow is that the specific design of a choice architecture matters because every difference in the space of options potentially makes a difference for people's choices. However, this does not give nudgers carte blanche to interfere with choice architecture. On the contrary, the inevitability of choice architecture rather *increases* the responsibility of policy makers *precisely because* choice architecture always influences

people in some way, regardless of whether it is the result of deliberate design or coincidence. According to the nudgers' own premise, choice architecture influences people either way, and thus policy makers are responsible even in case they decline to interfere with it in some area. If anything, then, the inevitability of choice architecture entails that there is a burden of justification for both interference and non-interference with choice architecture. In the context of government regulation, for example, the inevitability of choice architecture means that every governmental action—as well as every inaction—has consequences for people's lives and must therefore be justified. Whether a particular nudging practice is defensible, on the other hand, is a further question requiring careful ethical assessment of the characteristic features and the specific circumstances of its implementation.

I will now turn briefly to the main argument against nudging. The main ethical concerns critics have with nudging revolves around autonomy [21]. Broadly speaking, threats to autonomy in the context of paternalism can arise in two ways. Either paternalism interferes with the ends people set for themselves, or it interferes with the means people employ to achieve their ends, which is reflected in the common distinction between ends paternalism and means paternalism [9] (p. 19). Ends paternalism is seen as problematic because it imposes ends on people that are not necessarily their own. This leads to a mismatch between people's choices and what they actually want, which undermines their autonomy. When people are nudged in a certain direction, strictly speaking the choices they make are not their own, their "actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives" [18] (p. 128).

In reply, many libertarian paternalists claim that, contrary to standard forms of paternalism, nudging represents a form of the weaker means paternalism because it neither imposes ends on people nor questions the ends people have. Rather, nudging is intended to help people realize the ends they already set for themselves [10] (p. 433). Libertarian paternalists thus turn the tables on their critics and argue that nudging even promotes people's autonomy. For example, the autonomy of someone who wants to eat healthy but is tempted by less healthy options is actually enhanced, and not impaired, because the nudge only helped achieve an end the agent had anyways. Since nudges must be easy to avoid by definition, it is unlikely they would lead to changes of mind in staunch meat-eaters, which libertarian paternalists agree would be problematic because it would interfere with people's ends. In any case, they argue, autonomy is either preserved (no one is coerced, and freedom of choice is secured) or even promoted (the weak-willed are supported in realizing their ends).

However, means paternalism has also been met with criticism. Even if people's ends are respected, interfering with the means people employ to achieve them nevertheless represents a form of manipulation by "*bypassing their capacity for reason*" [22] (p. 5), which again is seen as undermining autonomy. As elaborated on in Section 2, nudges primarily target System 1 and thus exploit psychological vulnerabilities and faulty reasoning, such as inertia and framing. In doing so, critics argue, choice architects would not take people seriously as rational agents. Rather, they would take advantage of their cognitive weaknesses, even if they did so for their own benefit. For choices to be genuinely autonomous, critics insist, people not only have to be in control of setting their ends but also in control of the means and processes to realize those ends ([18] (p. 128)), ([23] (p. 209)).

In reply, some proponents of nudging complain that being fully in control of both ends *and* means represented too high a bar for choices to be autonomous and relied on an implausible conception of rational agency ([7] (p. 337)), ([21] (pp. 145–146)). Since choice architecture is inevitable, people are influenced one way or the other anyways, regardless of whether they are nudged or not. It would be an illusion to think that there could be purely rational processes free of any external influences. Thus, proponents argue, as long as people's ends are promoted or at least left intact, interfering with decision-making processes does not undermine autonomy.

The arguments for and against nudging are still subject to ongoing debate. As noted, in order to decide whether a given nudging practice is ethically problematic, its specific

features and the circumstances of its implementation must be taken into account. In the following section, I will therefore present relevant facts about climate change and the specific ethical challenges they raise, after which I will consider whether the objections from autonomy also apply to AI-powered nudges intended to induce more climate-friendly behavior.

4. Climate Change as an Ethical Challenge

Over the past decades, overwhelming scientific evidence has been gathered to substantiate the thesis that the current climate change—in particular global warming, rising sea levels, and increased frequency of extreme weather events—is caused primarily by anthropogenic GHG emissions [16]. Emitting GHGs of such magnitude has grave and long-lasting effects on the global climate system and will increase the probability of "irreversible impacts for people and ecosystems" [24] (p. 8). Negative effects of climate change outweigh the positive effects by far. Adaptation measures—adjustments to the adverse effects of climate change—are necessary in any case, but without substantial mitigation efforts—GHG emission reduction—the risk of harmful effects of more frequent and more intense climate and weather events will rise significantly [24] (pp. 18–19).

What is distinctive about the atmosphere is that it "comes closest to being a pure public good in that GHGs released anywhere have similar effects, making it a common as well as an essential resource" [25] (p. 79). While an essential resource, it is also finite in that its use without harmful environmental consequences is limited. Many ethicists thus consider climate change primarily a problem of justice, in particular a problem of intergenerational justice and distributive justice [25–27]. In essence, climate change as an ethical problem of intergenerational and distributive justice revolves around the question what present generations owe future generations [28] (chapter 1). The intergenerational aspect is due to climate change being a time-delayed phenomenon in that its effects extend far into the future because most GHGs have a very long lifetime in the atmosphere [24] (p. 87). Therefore, climate policy today will inevitably affect future generations. The distributive aspect is due to the fact that climate change brings with it an unequal distribution of burdens and benefits, which immediately raises questions such as how the cost of mitigation policies and the rights to emit GHGs are to be distributed fairly. For example, for a two-thirds chance of limiting global warming to 1.5 °C by 2050 the remaining carbon budget is roughly 420 $GtCO_2$ [29] (p. 12). If this goal is to be reached, then the atmosphere becomes a finite resource raising the question of just allocation of rights to use it.

The complexity of climate change forms a unique ethical challenge. Two aspects, in particular, pose extraordinary difficulties for an adequate response. First, the response necessary to fight climate change can be understood as a collective action problem [30]. Collective action problems often involve a multitude of agents who have an interest in using collective resources but are disincentivized to pay their fair share because of the possibility to benefit from the resources without carrying any burdens (what is also called "free riding"). Applied to the problem of climate change, this means all agents have an interest in using the atmosphere—through emission of GHGs—but individually they are disincentivized to contribute to the costs because from an individual perspective it is in their interest to free ride on the emission reduction efforts of others. Therefore, responses to climate change are often considered a variant of the tragedy of the commons, more specifically a prisoner's dilemma with a collective resource, in which it is collectively rational to reduce GHG emissions but not individually so [31] (p. 89). Second, GHG emissions are "externalities and are the biggest market failure the world has seen" [32] (p. 39). The costs of emissions in the form of harmful climate change are not fully paid by those who are causing them, but rather transferred to future generations. Thaler and Sunstein concede that in the face of market failures even libertarians think some form of government intervention may prove necessary [8] (p. 184). Examples are taxes on GHG emissions or cap-and-trade systems [8] (pp. 185–188). However, although Thaler and Sunstein do not reject such incentive-based approaches, they believe this approach should be supported with a nudging practice to reduce emissions because incentives like

taxes are often unpopular and therefore difficult for policy makers to implement. Instead, they argue, the hidden costs of emissions should be made visible to nudge people into action. To this end, they proposed that the government should devise a "Greenhouse Gas Inventory" in which the emissions by the biggest emitters are documented. This is supposed to raise public awareness, increase the pressure to act, and lead to more emissions reduction efforts [8] (p. 191).

The need for legal regulation becomes particularly apparent in view of the fact that large portions of global GHG emissions can be attributed to a comparatively small number of industrial companies, on which citizens have only limited influence [33]. Critics of nudging sometimes suggest that nudging practices are a bad substitute for structural reform, but proponents point out that there is no reason why we cannot do both [17]. I would thus agree with Sunstein and Thaler that market failures such as GHG emissions have to be primarily addressed by appropriate legislation but can also be accompanied by suitable nudging practices provided they are implemented in an ethically responsible way.

5. AI-Powered Climate Nudging

Against this backdrop, and given the opportunities of AI technology, one may ask: why not use green nudging as proposed by Thaler and Sunstein and combine it with AI in the effort to reduce GHG emissions? AI technology is already being used to fight climate change. For example, AI for Good, a non-profit organization, promotes using AI to advance the UN's Sustainable Development Goals (SDGs), among which is also "Climate Action" (SDG 13) [34]. In addition, Capgemini, a research institute, found that AI can contribute to combatting climate change when employed, for example, by companies to reduce emissions, improve energy efficiency, and optimize waste management [35].

In a recent book, Mark Coeckelbergh devises a thought experiment and envisions a "green brave new world" in which climate-related policy decisions are delegated to an AI-powered "Green Leviathan" because humanity did not manage to fight climate change by itself [15] (pp. 1–2). This, of course, represents a highly undesirable Huxleyan dystopia. Yet I think Coeckelbergh is perfectly right in highlighting the underlying problem he deals with in his book, namely "the problem of freedom in the light of climate change and AI" [15] (p. 5). In this context, he also explores AI-powered climate nudging, a seemingly freedom-preserving alternative to the authoritarian Green Leviathan:

> One could imagine that nudging is used for changing individual behavior in a more environmentally and climate-friendly direction. [. . .] And maybe AI, having analyzed the data of entire populations or even the entire world, could give us statistical information about our collective carbon footprints and communicate this information in a way that has similar effects on us. Not just by providing information as such, and not by persuasion by means of rational arguments, but by working with human biases and emotions. In this way, nobody is forced to do the right thing, as an authoritarian regime would do; instead, people are 'gently' pushed in a direction. But they can always opt out, they can always make other choices. It seems that freedom is preserved. [15] (pp. 36–37)

Eventually, however, Coeckelbergh rejects AI-powered climate nudging as a form of paternalism infantilizing people because the government would treat its citizens as irrational children incapable of doing the right thing. Although libertarian paternalists insisted on promoting people's own ends "in practice someone else judges for them: the nudger" [15] (p. 38). Coeckelbergh further argues that exploiting cognitive weaknesses–System 1 vulnerabilities—represents a form of unacceptable manipulation undermining people's autonomy by disrespecting their rational capacities [15] (p. 39). He concludes that the tactics employed by choice architects revealed a profound distrust in people since they operated on the assumption "that humans are weak-willed or irrational, and do not always know what is good for them" [15] (p. 41).

Coeckelbergh's critique echoes some aspects of the arguments against nudging touched on in Section 3. Additionally, the employment of AI technology to nudge people into more

climate-friendly behavior exacerbates the ethical concern with autonomy because of the large-scale effects it may produce. Arranging food items in a cafeteria is a nudging practice with quite local effects, but in a digital sphere with millions of users, nudges could have a rather pervasive impact. For this scenario to be plausible we do not need to imagine a Green Leviathan. There is already evidence how character traits can be derived from digital footprints for effective mass persuasion [36]. For example, just think of the infamous Facebook experiment in which the news feeds of almost 690,000 people were manipulated [37].

The potentially large-scale effects of AI-powered nudging add the societal dimension to threats against autonomy. Not only the autonomy of specific (groups of) individuals is endangered, but also the autonomy of society as a whole, its collective autonomy to determine societal ends and the means to pursue them. This means AI-powered nudging may pose a threat to democracy itself. Objections against nudging based on a concern for individual autonomy can thus also be raised out of worries about collective autonomy since in both cases the self-determination of ends and means is interfered with. In Section 3 I elaborated on two ways in which individual autonomy can be undermined, either by interfering with people's ends (ends paternalism) or with the means people employ to realize their already existing ends (means paternalism). These two ways of violating people's autonomy can also occur on the societal level. With regard to ends paternalism, critics fear that a "data-empowered 'wise king'" would be in a position "to produce desired economic and social outcomes almost as if with a digital magic wand", which could ultimately lead to a "top-down controlled society" [14]. In this scenario, people would no longer govern themselves through democratic processes but rather be controlled by a quasi-totalitarian regime imposing its ends on the citizenry. With regard to means paternalism, critics have argued that just as individual autonomy is undermined by exploiting cognitive weaknesses because it bypasses people's capacity for reason, in the same way the collective autonomy of a democratic society is undermined if choice architects "bypass public debate and opt for psychological manipulation instead" [38]. In this scenario, even if people's ends are respected they are still manipulated because hidden influence is exerted to interfere with the means they use for achieving their ends.

These are all legitimate concerns and objections, but do they also apply to the case of climate nudging? As emphasized in Section 2, the specific ends, means and justifications involved in a particular nudging practice must be taken into account to determine its ethical permissibility. In addition, there seem to be certain differences between climate nudging and more standard or conventional forms of nudging such as the cafeteria example.

First, the distinctive characteristic of paternalistic nudges consists in their aim to increase an agent's individual welfare. Yet this is not the case where nudges aim at getting people to behave in more climate-friendly ways. Rather, climate nudging aims at protecting the environment and future generations from harm caused by excessive GHG emissions generated in the present. As pointed out, climate change is a time-delayed phenomenon, therefore cutting emissions now would particularly benefit future generations and not (only) the people at which the nudges contributing to the reduction are aimed. As I also pointed out, emissions are an externality and thus represent a type of market failure. Drawing on the distinction in Section 2, climate nudges—just as green nudges in general [7] (p. 331)—can be categorized as market failure nudges rather than paternalistic nudges.

Second, the difference regarding the aim of climate nudging (protecting the environment and future generations) compared to paternalistic nudging (improving individual welfare) also opens up the possibility of a different justification. What critics of paternalism generally take issue with is that the interference is purported to be justified with reference to the presumption of a third party to know better what is good for individuals than individuals themselves. This presumption is indeed problematic because in order to know what is best for an individual a third party would have to know the individual's personal preferences. However, who is in a better position to know one's personal preferences than oneself? That is why many critics of paternalism draw on Mill's famous no-harm principle, according to which the only condition under which interference with people's

liberty is legitimate "is to prevent harm to others. His own good, either physical or moral, is not sufficient warrant" [39] (p. 80). However, as pointed out in the previous section, the atmosphere is an inherently public good. Determining whether certain activities harm the environment in the form of adverse effects of climate change does not necessarily require taking into account the personal preferences of particular individuals, or at least they are less relevant. Rather, one needs to look primarily at scientific research, and here the jury is in: overuse of the atmosphere's capacity to absorb emissions will have harmful effects for the environment and future generations [16]. It is therefore not uncommon in the debate over moral obligations regarding climate change to appeal to the no-harm principle [40] (p. 218). Accordingly, it could also be used to justify climate nudging because it would be done to protect others from harm—future generations—and not because the government presumes to know what is better for particular individuals.

The third ethical issue concerns the means employed by climate nudging. Assuming that protecting the environment and future generations from harm is a legitimate end, would climate nudging then be justified? Here I agree with critics that for choices to be genuinely autonomous a mere focus on ends is insufficient. In my view, a central discomfort underlying ethical qualms about paternalism in general and nudging in particular is the structural asymmetry between nudgers and nudgees. *Someone else* occupies an allegedly superior vantage point and interferes with one's choices and decision-making processes. Even if this third party had our best interests at heart, exploiting our cognitive weaknesses to further our interests would still disrespect autonomy. While it may be true that we should have a realistic understanding of the inner workings of our cognitive functions and acknowledge that our choices can also be influenced on a subconscious level in some way, exploiting these influences with non-transparent manipulation techniques seems not the right way to deal with those weaknesses.

Given all these considerations, is it possible to implement climate nudging practices that avoid undermining autonomy and democracy? I think climate nudging can be ethically permissible if it takes the form of *self-governance*. Nudging as self-governance comprises at least the following three conditions, which have to be met in order for a climate nudging practice implemented by policy makers to be ethically permissible. These conditions are best thought of as interrelated aspects of a single overall ethical assessment rather than isolated boxes that could be checked completely independently from one another:

- *Symmetry Condition*: nudgers and nudgees should be at least structurally identical, that is, those groups of individuals or their representatives initiating or approving a particular nudging practice should be the same groups of individuals potentially affected by this practice.
- *Democracy Condition*: a policy implementing a particular nudging practice should possess a democratic mandate in some form, that is, the implementation of a nudging practice should require a procedure of public debate and approval.
- *Transparency Condition*: a particular nudging practice should be implemented in a way so that, in principle, everyone can identify the practice and learn about its mechanisms.

The point of a nudging practice taking the form of self-governance that satisfies these three conditions is that a society implementing such a nudging practice would effectively nudge itself in a self-determined, democratically legitimate, and transparent way. Ethical problems associated with asymmetry and manipulation can be avoided because people's ends are respected and their means are not exploited. Take, for example, the cafeteria scenario again. Even if people's ends are respected and promoted, some critics still object to putting healthy food items at eye level because it interfered with people's decision-making process in a non-transparent and therefore manipulative way. But imagine the cafeteria was a school cafeteria and all parties involved—for example, students, parents, teachers, etc.—decided together to make the menu healthier and implemented in a transparent way a nudging practice and accompanied it with an information campaign. I think in this case there would be much less occasion for ethical concerns. In the following, I present and discuss an example of a green nudging practice and of an AI-powered climate nudging

practice, from the public sector and from the private sector, respectively, that may serve to illustrate what I have called self-governance.

In the late 1980s, an environmental initiative was founded in the German town Schönau [41]. It proposed to take over the local power grid and energy provider, and after campaigns and public debate they put it to a vote. The proposal was accepted, the energy provider became the standard utility and adopted a "green default", meaning most of the energy comes from renewable sources. As a consequence, customers are provided with green energy unless they opt out. By 2006, almost all Schönau households used green energy.

In this example, a default—a standard nudging tool—was implemented after public debate and a subsequent vote, thus satisfying all three conditions I characterized above. Many critics of nudging take issue with defaults because they make use of a System 1 psychological vulnerability (inertia) and thus represent a form of manipulation. However, I think this is not the case in this example. For one thing, the nudge is transparent and does not operate behind people's back. Assuming that people do in fact stick to a non-green energy default out of inertia, in the Schönau case this System 1 psychological vulnerability was not dealt with as a weakness to be exploited surreptitiously, but rather as an issue to be addressed openly. For another, the nudge was democratically implemented by the people potentially affected so that there is no asymmetry between nudgers and nudgees, as would have been the case if a government had implemented the policy in a top-down manner. Finally, if you oppose the green default, for whatever reason, you still can opt out, so freedom of choice is also preserved. As a result, the citizens of Schönau basically nudge themselves in a self-determined, democratically legitimate, and transparent way.

A concrete example for AI-powered climate nudging from the private sector is a service offered by Google. Google now provides users of Google Flights with carbon emissions information so users can include the carbon footprint of different flights into their decision-making process and thus may consciously choose alternatives with lower emissions [42]. Search results for flights display in a prominent way estimates of how much kg of CO_2 is specific for a particular flight. Additionally, flights are marked with a "green badge" if they are associated with much less emissions compared to the amount of emissions typical for this route, and also display the amount of emissions saved by choosing this alternative. A further feature is that if for a particular route a train connection is available, then the train connection will also be listed among the results together with the carbon emissions information.

In my view, including carbon emissions information in the search results for flights constitutes a nudging practice, in particular the "green badge" marking flights with significantly less emissions. This type of providing information relevant for climate-friendly (or at least climate-friendlier) behavior can be considered an example of eco-labelling, which in turn is a staple of green nudging [7] (p. 332). One of the mechanisms underlying eco-labelling is the so-called "salience bias", which can be operative in real-world as well as digital environments [43] (p. 7). According to the salience bias, people tend to focus on aspects of their environment that stand out in some way or other. The classic cafeteria scenario in which healthy food items are made salient by putting them at eye level is a case in point. Likewise, making search results salient by marking them with a "green badge" thus seems to qualify as a green nudge, in particular a climate nudge, because this is intended to steer people's choices towards more climate-friendly options.

Now, does this nudging practice satisfy the three conditions I characterized above, and can it therefore be considered an ethically permissible nudging practice assuming the form of self-governance? First of all, since the nudging practice is not part of a policymaking tool but a service from the private sector, the democracy condition does not really apply. However, consumers are free to use the service or not so there need not be a public procedure of approval. Yet this also means that the satisfaction of the other two conditions is even more important. The symmetry condition requires that consumers not be manipulated by a third party, but rather knowingly and voluntarily nudge themselves into climate-friendly behavior by using the service. Here one can see how the conditions are interrelated, because

for the symmetry condition to be fulfilled the transparency condition must simultaneously be fulfilled. I can only nudge myself if I know that I am participating in a corresponding practice. This seems to be the case here because the service is transparent about carbon emissions information being a relevant factor in displaying the search results.

Nevertheless, some critics of nudging argue that exploiting the salience bias is manipulative on the grounds that the salience of an item and its actual importance may come apart. Taking this critique into account, Robert Noggle argues that "a salience nudge is not manipulative if it influences choice by bringing the salience of some fact into closer alignment with its actual importance" [44] (p. 168). Of course, it is often difficult to determine the importance of some fact, in particular when people's personal preferences are involved. Again, as I argued above, what the nudging practice in this case aims at is not the personal welfare of individual agents but an inherently public good—the atmosphere. Additionally, there is no doubt that the carbon emissions produced by air travel contribute to climate change. Making carbon emissions salient thus brings them in alignment with the important role they play in contributing to climate change. Ethicists often point out that one of the problems in dealing with climate change consists in insufficient awareness of the harmful consequences of activities involving emissions because they are quite literally invisible, and their adverse effects are distributed spatially and temporally [31] (p. 88). Drawing attention to these effects and making them visible to raise awareness about the consequences of our actions seems not to be manipulative. Since there is no restriction on the set of options, freedom of choice also seems to be preserved, and it is always possible to opt out of the service entirely.

Of course, what is primarily necessary is structural reform aiming at decarbonization. However, the fight against the harmful consequences of climate change also faces the problem of "institutional inadequacy" [31] (p. 89) because enforceable sanctions required for limiting GHG emissions are difficult to implement on a global level. Thus, to reduce carbon emissions there seems to be no reason why we should not also engage in effective and ethically permissible climate nudging while pursuing structural reform. In sum, if I nudge myself into taking the train instead of a domestic flight because I am made aware in a non-manipulative way of the significant amount of emissions a flight for the same route would cause, then this seems to be a genuinely autonomous choice and not a case of problematic manipulation.

6. Conclusions

Whereas many nudging practices are in fact ethically problematic, I maintain AI-powered climate nudging can be ethically permissible if it takes the form of self-governance satisfying the symmetry condition, the democracy condition, and the transparency condition. A society implementing corresponding policies would nudge itself and therefore avoid the asymmetry between nudgers and nudgees as well as the danger of manipulation asymmetry involves. Of course, the Green Leviathan is definitively not a role model for solving the climate crisis with AI technology. The harm to autonomy and democracy would represent an unacceptable damage, even if motivated by good intentions. The gold standard of policymaking should always be rational persuasion, and information campaigns should not be replaced by nudging practices. However, if the ethical assessment of a particular climate nudging practice powered by AI takes the form of self-governance, then the power of AI can be harnessed in an ethically justifiable way as a supporting measure to fight the adverse effects of climate change.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jacques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *arXiv* **2019**, arXiv:1906.05433. [CrossRef]
2. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
3. van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
4. Coeckelbergh, M. AI for climate: Freedom, justice, and other ethical and political challenges. *AI Ethics* **2021**, *1*, 67–72. [CrossRef]
5. Floridi, F.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 698–707. [CrossRef] [PubMed]
6. van Wynsberghe, A. *Artificial Intelligence. From Ethics to Policy*; Study, Panel for the Future of Science and Technology, European Parliamentary Research Service (EPRS), Scientific Foresight Unit (STOA); European Union: Brussels, Belgium, 2020. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641507/EPRS_STU(2020)641507_EN.pdf (accessed on 7 April 2022).
7. Schubert, C. Green Nudges: Do they work? Are they ethical? *Ecol. Econ.* **2017**, *132*, 329–342. [CrossRef]
8. Thaler, R.H.; Sunstein, C.R. *Nudge. Improving Decisions about Health Wealth, and Happiness*; Yale University Press: New Haven, CT, USA; London, UK, 2008.
9. Sunstein, C.R. *Why Nudge? The Politics of Libertarian Paternalism*; Yale University Press: New Haven, CT, USA; London, UK, 2014.
10. Sunstein, C.R. The Ethics of Nudging. *Yale J. Regul.* **2015**, *32*, 413–450. [CrossRef]
11. Weinmann, M.; Schneider, C.; vom Brocke, J. Digital Nudging. *Bus. Inf. Syst. Eng.* **2016**, *58*, 433–436. [CrossRef]
12. Mirsch, T.; Lehrer, C.; Jung, R. Digital Nudging: Altering User Behavior in Digital Environments. In Proceedings of the 13th International Conference on Wirtschaftsinformatik, St. Gallen, Switzerland, 12–15 February 2017; pp. 634–648.
13. Yeung, K. 'Hypernudge': Big Data as a Mode of Regulation by design. *Inf. Commun. Soc.* **2017**, *20*, 118–136. [CrossRef]
14. Helbing, D.; Frey, B.S.; Gigerenzer, G.; Hafen, E.; Hagner, M.; Hofstetter, Y.; van den Hoven, J.; Zicari, R.V.; Zwitter, A. *Will Democracy Survive Big Data and Artificial Intelligence?* Scientific American: New York, NY, USA, 2017. Available online: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/?redirect=1 (accessed on 7 April 2022).
15. Coeckelbergh, M. *Green Leviathan or the Poetics of Political Liberty*; Routledge: New York, NY, USA; London, UK, 2021.
16. IPCC. *Climate Change 2021: The Physical Science Basis*; IPCC: Geneva, Switzerland, 2021. Available online: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf (accessed on 7 April 2022).
17. Schmidt, A.T.; Engelen, B. The ethics of nudging: An overview. *Philos. Compass* **2020**, *15*, e12658. [CrossRef]
18. Hausman, D.M.; Welch, B. Debate: To Nudge or Not to Nudge. *J. Political Philos.* **2010**, *18*, 123–136. [CrossRef]
19. Dworkin, G. Paternalism. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Stanford, CA, USA, 2020. Available online: https://plato.stanford.edu/archives/fall2020/entries/paternalism/ (accessed on 7 April 2022).
20. Kahneman, D. *Thinking, Fast and Slow*; Penguin: London, UK, 2012.
21. Engelen, B.; Nys, T. Nudging and Autonomy: Analyzing and Alleviating the Worries. *Rev. Philos. Psychol.* **2020**, *11*, 137–156. [CrossRef]
22. Blumenthal-Barby, J.S.; Burroughs, S. Seeking Better Health Care Outcomes: The Ethics of Using the "Nudge". *Am. J. Bioeth.* **2012**, *12*, 1–10. [CrossRef]
23. Bovens, L. The Ethics of Nudge. In *Preference Change. Approaches from Philosophy, Economics and Psychology*; Grüne-Yanoff, T., Hansson, S.O., Eds.; Springer: Dordrecht, The Netherlands, 2009; pp. 207–219. [CrossRef]
24. IPCC. *Climate Change 2014: Synthesis Report*; IPCC: Geneva, Switzerland, 2014. Available online: https://www.ipcc.ch/site/assets/uploads/2018/02/SYR_AR5_FINAL_full.pdf (accessed on 7 April 2022).
25. Vanderheiden, S. *Atmospheric Justice*; Oxford University Press: Oxford, UK, 2008.
26. Page, E.A. *Climate Change, Justice and Future Generations*; Edward Elgar: Cheltenham, UK; Northampton, UK, 2006.
27. Caney, S. Climate Change. In *The Oxford Handbook of Distributive Justice*; Olsaretti, S., Ed.; Oxford University Press: Oxford, UK, 2018; pp. 664–688. [CrossRef]
28. Hiskes, R.P. *The Human Right to a Green Future: Environmental Rights and Intergenerational Justice*; Cambridge University Press: Cambridge, UK, 2009.
29. IPCC. *Global Warming of 1.5 °C*; IPCC: Geneva, Switzerland, 2018. Available online: https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_Low_Res.pdf (accessed on 7 April 2022).
30. Hayward, T. Climate change and ethics. *Nat. Clim. Chang.* **2012**, *2*, 843–848. [CrossRef]
31. Gardiner, S.M. A Perfect Moral Storm: Climate Change, Intergenerational Ethics, and the Problem of Moral Corruption. In *Climate Ethics. Essential Readings*; Gardiner, S.M., Caney, S., Jamieson, D., Shue, H., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 87–98.
32. Stern, N. The Economics of Climate Change. In *Climate Ethics. Essential Readings*; Gardiner, S.M., Caney, S., Jamieson, D., Shue, H., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 39–76.
33. CDP Carbon Majors Report. 2017. Available online: https://cdn.cdp.net/cdp-production/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf?1501833772 (accessed on 7 April 2022).

34. AI for Good. Available online: https://ai4good.org/ai-for-sdgs/goal-13-climate-action/ (accessed on 7 April 2022).
35. Capgemini Research Institute. Climate AI: How Artificial Intelligence Can Power Your Climate Action Strategy. 2020. Available online: https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2020/11/Report_Climate_AI_Capgemini_Research_Institute.pdf (accessed on 7 April 2022).
36. Matz, S.C.; Netzer, O. Using Big Data as a window into consumers' psychology. *Curr. Opin. Behav. Sci.* **2017**, *18*, 7–12. [CrossRef]
37. Kramer, A.D.I.; Guillory, J.E.; Hancock, J.T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8788–8790. [CrossRef] [PubMed]
38. Furedi, F. Defending Moral Autonomy against an Army of Nudgers. Spiked. 2018. Available online: https://www.spiked-online.com/2011/01/20/defending-moral-autonomy-against-an-army-of-nudgers (accessed on 7 April 2022).
39. Mill, J.S. *On Liberty*; Yale University Press: New Haven, CT, USA; London, UK, 2003.
40. Singer, P. Climate Change. In *Practical Ethics*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2011; pp. 216–237.
41. Pichert, D.; Katsikopoulos, K.V. Green defaults: Information presentation and pro-environmental behaviour. *J. Environ. Psychol.* **2008**, *28*, 63–73. [CrossRef]
42. Find Flights with Lower Carbon Emissions. Available online: https://blog.google/products/travel/find-flights-with-lower-carbon-emissions/ (accessed on 7 April 2022).
43. Caraban, A.; Karapanos, E.; Gonçalves, D.; Campos, P. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–15. [CrossRef]
44. Noggle, R. Manipulation, salience, and nudges. *Bioethics* **2018**, *32*, 164–170. [CrossRef] [PubMed]