# Democracy and Paternalism

*Axel Ockenfels & Carl Christian von Weizsäcker*

## A. Introduction

In Wikipedia we read the following definition of paternalism:

> **"Paternalism** is action that limits a person's or group's liberty or autonomy against their will and is intended to promote their own good. It has been defended in a variety of contexts as a means of protecting individuals from significant harm, supporting long-term autonomy, or promoting moral or psychological well-being. Such justifications are commonly found in public health policy, legal theory, medical ethics, and behavioral economics, where limited intervention is viewed as compatible with or even supportive of personal agency."

This definition is rooted in John Stuart Mill´s book "On Liberty" 1859/1991. So, the Wikipedia entry continues, quoting Mill:

> "It is, perhaps, hardly necessary to say that this doctrine [i.e. that individual liberty should only be restricted to protect a person or to protect others] is meant to apply only to human beings in the maturity of their faculties. We are not speaking of children, or of young persons below the age which the law may fix as that of manhood or womanhood."

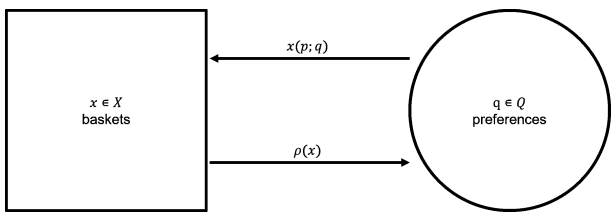
Paternalism then is an attitude which promotes and justifies action from "above" to control and guide actions of others considered to be "below" those who are their "guides". The idea of "Democracy" corresponds to the opposite: the people "below" choose their agents, who are supposed to solve any joint problems in the interest of the people. Democracy, well understood, is a form of government of people who are free: liberty is a precondition of democracy.

Based on earlier work by von Weizsäcker 2024 and Ambuehl, Bernheim, Ockenfels 2021 we try to come to grips about the relation between the two contrarian concepts of "paternalism" and "democracy".

## B. Preferences and Freedom

Weizsäcker 2024 is about a consistent concept of welfare economics for the case that preferences of citizens are endogenously determined. The book generalizes traditional welfare economics, as pioneered by Samuelson 1938. In traditional welfare economics preferences are assumed to be exogenously given. In Weizsäcker 2024 it is shown that this welfare economic approach can be generalized, if and only if endogenously formed preferences are "adaptive". Further, the book also shows that the results of behavioral economics are consistent with the hypothesis of "adaptive preferences".

In the Weizsäcker 2024 approach endogenous preferences are modeled by means of the concept of a "preference system". We assume the existence of a space $\mathfrak{X}$ of potential consumption baskets, and we assume the existence of a space $\mathbb{Q}$ of potential preferences. The following figure indicates the idea of a preference system.




There is a mapping $x(p;q)$ from the space of preferences $\mathbb{Q}$ into the space of consumption baskets. It is the well-known demand function: the price vector $p$ indicates the budget constraint; the preference vector $q$ is an element of the space of preferences $\mathbb{Q}$. What distinguishes the preference system from the conventional demand function is the additional presence of a mapping $q = \rho\,(x)$ from the space of consumption baskets $\mathfrak{X}$ to the space of possible preferences $\mathbb{Q}$. This makes preferences an endogenous variable. Here we assume that it is the actual consumption vector $x$ which, with some time delay, induces preferences. This is the concept of "induced preferences".

Within this general framework of a preference system, we then introduce a special case: it is the concept of "adaptive preferences". Intuitively speaking, adaptive preferences exhibit the feature of a certain preference

conservatism. To take as an example the well-known phenomenon of loss aversion. Within a conventional framework of the von-Neumann-Morgenstern utility function, loss aversion violates the assumption of fixed preferences. However, loss aversion exhibits a certain preference conservatism: people put an extra weight on a situation which avoids losses of their *actual* level of wealth.

The formal definition of adaptive preferences is this: consider two consumption baskets $x$ and $y$ such that with preferences $q = \rho(x)$ induced by $x$ basket $y$ is preferred over $x$. Preferences are defined to be *adaptive*, if it then is also the case that $y$ is preferred over $x$, if preferences are induced by $y$.

By introducing the idea of "progress" or "improvement", we can provide an intuition, why adaptiveness of preferences is an important ingredient for welfare economics in a world of endogenously formed preferences. For simplicity of presentation, assume that inducement of preferences happens with a delay of one year. Consider the following sequence of baskets through time. In the first year we have basket $x$ which induces preferences $q = \rho(x)$. These preferences prevail in the second year. Now, in the second year we have consumption basket $y$ of which we assume that with prevailing preferences $q = \rho(x)$ the basket $y$ is preferred over $x$. So, the move from $x$ to $y$ can be called an improvement. Let us then assume that in year 3 the basket $x$ again prevails. Preferences in year 3 are induced by the basket $y$, which prevailed in year 2. Assume now that, contrary to the assumption of adaptive preferences, with preferences induced by $y$ basket $x$ is preferred over basket $y$. We then observe an improvement sequence from $x$ to $y$ and then back to $x$. We call this improvement sequence a pseudo-improvement. It is obvious that welfare economics basically becomes impossible, if such pseudo-improvements cannot be ruled out.

The assumption of adaptive preferences makes this particular kind of pseudo-improvements impossible. A large part of Weizsäcker 2024 is devoted to the proof that within the models developed in the book the assumption of adaptive preferences is sufficient to rule out pseudo-improvements and that, thereby, welfare economics of endogenously determined preferences is possible.

Theorems 1 and 2 provide such proof. Generally, we assume that the mappings between $\mathfrak{X}$ and $\mathbb{Q}$ are continuous for the topologies of these spaces. Furthermore, we have a non-satiation assumption and a certain single-crossing assumption. We operate in a discrete time model and in

addition in a continuous-time model. The latter we call the "real world model", because the real world exhibits continuous time.

Theorem 1 (in its different varieties, depending on the specifics of different models) says that, with adaptive preferences, any sequence of improving baskets of any finite length is a-cyclic, i.e., cannot come back to the origin.

Theorem 2 (in its different varieties) says this: if every sequence of improving baskets of any finite length is acyclic then preferences are adaptive.

There is then equivalence between the assumption of adaptive preferences and the assumption that welfare economics remains possible, when preferences are endogenously formed. As explained in detail in the book, for the case of endogenously formed preferences, the assumption of adaptive preferences can be seen as the generalization of the Samuelson-Houthakker axioms of revealed preference, which apply for the traditional model of fixed preferences. Note that fixed preferences are a special case of adaptive preferences.

As discussed in the book, not everybody in society acts in an autonomous mode. Thus, for example, small children do not behave in their own best interest. Every society accepts that parents interfere with their own small children so as to avoid behavior which is deleterious to the actor himself/herself. Freedom or liberty reasonably presupposes that people behave in an autonomous manner. However, autonomous behavior is not innate but must be acquired. This is the reason that every society requires a certain minimum age before people obtain all the rights a society of free people provides. Weizsäcker 2024 discusses this autonomy requirement for a society of free people (for example, Chapters 21, 22, 23, 24, 27).

The central idea for the path towards autonomy is "education". Successful schooling and education always requires that preferences are adaptive (Weizsäcker 2024, Chapter 20, Section 5). If we believe in education, we already implicitly believe in the prevalence of adaptive preferences. However, education is required for the goal of autonomous behavior. "Education" derives from Latin "educare" = educate, related to "educere" = "lead out" (Oxford Dictionary of English Etymology, Oxford 1966). By appropriate education people are brought (led out) from infantile immaturity towards autonomy.

As history has shown, such education towards civil liberty is not self-evidently successful. Karl Popper pointed to the great philosopher Plato as the intellectual fountain of 20th Century totalitarian ideology (Weizsäcker 2024, Chapter 27). In Plato´s "Republic" the "philosopher king" was the

only man in society with a successful "education" towards autonomy. And history, as well as our present world politics tell us of the many successes of such totalitarian ideology. There is the great temptation of solving the freedom problem by simply abolishing freedom.

Thus, in a sense, a democratic society of free people is not free in its choice of the details of the great educational enterprise toward citizens´ autonomy. It has to take account of the pitfalls of education towards mature citizenship. There are constraints to be taken in account in the design of educational procedures. In our present paper, we do not discuss the general characteristics of this educational enterprise. We only consider one aspect, which is personal intertemporal choice.

Individual intertemporal choice, for example in terms of consumption and labor supply, is very much linked to the subjectively perceived causation of future events. Animals have little understanding of causal chains. They simply follow their instincts, like hunger, fatigue, fear, curiosity, attachment to their mother, and, when grown up, sexual drives. Similarly, small children mainly follow their instincts. Paradise is the place, where their wants, generated by their instincts, are immediately satisfied. Paradise is the place with no need for intertemporal choice. Growing human understanding of the causal connection of events in the world at large, imposes on human action the need to anticipate the future. This leads to the need to think in terms of intertemporal trade-offs. It is the need to understand the opportunity cost of immediate satisfaction of wants. "Paradise lost". To learn about intertemporal trade-offs is an important part of the education towards autonomy.

The social process of education involves a majority of people as "educators", as "teachers". Parents are educators of their children; the social production machine is at the same time a complex educational machine with experienced workers as teachers and newcomers, apprentices as pupils. Modern society owns many schools, from kindergarten up to universities. We should understand that a large fraction of altruistic behavior within the family, among friends and colleagues takes the form of educating other people.

## C. Experimental Evidence on Paternalism

Educating others is a quintessential form of altruism, and indeed of paternalism. The central dilemma is the tension between helping others

to make better decisions, and promoting learning while respecting the learner's autonomy.

Recent experimental evidence by Ambuehl, Bernheim, and Ockenfels (2021) shows how ordinary "choice architects" (CAs) navigate this trade-off. In their laboratory market for intertemporal payments, CAs design opportunity sets for "choosers". They regularly delete impatient options, even when non-binding advice could have been offered instead, because they believe the choosers will be better off.

Thus, in practice, education often takes the form of restricting choice rather than providing information.

However, the experiment also reveals the challenges associated with such interventions: CAs lack reliable information about another individual's preferences. How, then, do they decide which paths are helpful and which are harmful? Ambuehl et al. trace their judgements to the projective inferences first described by Adam Smith in his Theory of Moral Sentiments. These inferences follow from self-examination: "As we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation." (Part I, Section I, Chapter I).

In this spirit, Ambuehl et al. identify two types of paternalism: those who reason about others based on their own mistakes and those who reason based on their own preferences.

A mistakes-projective paternalist assumes that others tend to share their susceptibility to error. They behave as if they are trying to help others avoid choices that they themselves would reject, but which they nevertheless are tempted to choose. This inclination generates a negative correlation between the choices she makes for herself and the restrictions she imposes on others in Ambuehl et al.'s setup.

In contrast, an ideals-projective paternalist behaves as if her own preferences are relevant to others, either because she believes they share her values or because she thinks her perspective is valid and theirs is not. Ideals-projective paternalism generates a positive correlation between the choices paternalists make for themselves and the restrictions they impose on others.

As ideals-projective reasoning very much dominates in the data, interventions are systematically misguided, even according to the CAs' own welfare criteria. They remove impatient options even when all payoffs are delayed, i.e. in settings that neutralize present bias arguments. In other words, CAs do not intervene by removing options that they wish they could re-

sist when choosing for themselves (mistakes-projective paternalism); rather, they intervene as if they seek to align others' choices with their own aspirations (ideal-projective paternalism). False consensus bias compounds the error: patient CAs underestimate how many choosers they influence.

Furthermore, Ambuehl et al. demonstrate that the relationships between policy preferences and consumption outside the laboratory are consistent with ideals-projective paternalism. For example, lighter drinkers express significantly more support for increasing alcohol taxes in another jurisdiction.

More recent evidence from Isoni, Ockenfels, Sugden and Zheng (2025, in progress) complements these findings with data from the demand side. They begin with the observation that political approval for "helping" paternalistic interventions (i.e. the supply of paternalism) and the observation that some people choose to self-constrain (e.g. Sunstein and Thaler's New Year's resolution test) are sometimes presented as evidence of demand for paternalism. However, the supply of paternalism by CAs may be motivated by a desire to help or be helped (or both). In other words, a preference for self-constraint does not necessarily imply consent for externally imposed constraints.

Indeed, in a series of experiments, Isoni et al. found that subjects reported believing that the paternalistic decisions of an unknown other would be similar to their own, and that their own decisions were very similar to their inner demands. However, many subjects were happy to bind themselves, but objected to being bound by an unknown other. The demand for paternalism is much smaller than the supply.

Isoni et al.'s data exclude some hypotheses for this pattern, such as the existence of naïve individuals who do not realize they need help, or that subjects value decision autonomy in itself. An alternative approach could be based on bounded cognition: intuitive heuristics suffice for self-constraint and other-constraint, but make it difficult to evaluate whether constraints imposed by others are desirable.

To summarize, Weizsäcker (2024) provides a normative basis for combining welfare analysis and behavioral economics. His preference system recognizes that consumption today shapes preferences tomorrow. Welfare economics is only possible when preferences are adaptive, i.e. when improvements to welfare under current tastes are not undone once tastes adapt. In this context, adaptive preferences may justify educative or even paternalistic guidance as a learning aid on the path to autonomy.

Laboratory experiments, on the other hand, reveal the psychological

mechanisms of education and paternalism. While well-meaning, people often overstep the mark when intervening in others' choices, eventually limiting autonomy in a way that decreases welfare. CAs rely on their own (adapted) ideals as the best available proxy for the learner's preferences. However, the false consensus bias makes CAs overconfident in the relevance of their own ideals to learners. Moreover, while people seem happy to restrict the autonomy of others, they may not fully consider that they would reject being paternalized themselves.

Educational paternalism need not be a contradiction. When adaptive preferences make some paternal guidance unavoidable, the ideals of projective educators can provide the social energy that drives progress – but they must be guided by evidence of actual learner welfare. Blending Weizsäcker's macro theory with micro evidence from Ambuehl et al. and Isoni et al. provides an explanation of the circumstances in which paternalism promotes, rather than diminishes, autonomy.

Of course, there are forms of education that do not require direct behavior control. Providing people with advice and more options to empower them to make informed decisions that respect and extend one's authority is often possible. Ockenfels (2023), among others, provides examples in "behavioral market design". There does not always need to be a conflict between autonomy and education.

## D. Institutions

There is a history of institutions spanning thousands of years. And, a long history of political philosophy reflecting the observed gradual or abrupt changes of prevailing institutions. What can we learn for our theme from this history? It is obvious that successful working of democracy requires an institutional framework. Historical attempts to abandon all existing institutions and to replace them with completely different ones have always failed. As examples we may point to the developments of the great French revolution, 1789 onwards, to the Russian revolution of 1917 onwards, or to Mao Zedong's "Cultural Revolution" in the sixties and seventies of the last century.

We then can ask: how much paternalism is contained in those institutions which are typical for flourishing modern democracies? Take the case of franchise or suffrage. Historically, suffrage was more restricted than it is today. In most countries suffrage for women was not introduced earlier

than in the 20th century. Obviously then the laws concerning franchise in earlier times contained a lot of paternalism.

Another, more subtle form of paternalism is the two-houses organization of legislatures. The Constitution of the United States of America prescribes that the legislature, the Congress, consist of two chambers: the Senate and the House of Representatives. Whereas there are general elections for the House every two years, senators are elected for a period of six years. The Constitution thereby expresses some distrust into the majority decisions of the House. Underlying is the well-known fact that the short-run effects of law-changes differ from their long-run effects. And the further idea that the voters do not fully understand the consequences of law changes. There is, of course, a substantial economics literature about the ensuing political short-termism.

The longer election period for senators then can be seen as a kind of constitutional paternalism – in this case against excessive short-termism of political decisions. We should add that the fathers of the American constitution were strongly influenced by Montesquieu, in particular by his great work on the "Spirit of Law". There he expounds on the "separation of powers" into legislative, executive, and judiciary.

The problem of potential short-termism is also at the core of the institution of an independent central bank. After the disastrous experience of two German hyperinflations after World War I and after World War II, the Federal Republic of Germany decided in the early fifties of the 20th century that the newly established central bank should obtain a high degree of independence. The Bundesbank then was able to make the Deutsche Mark a rather stable currency. In his work across several decades Alex Cukierman has empirically well established that currencies under control of an independent central bank had a substantially smaller inflationary bias than currencies under a more direct control of democratically elected governments: cf. for example Cukierman 1992. There is a short-termism of democratically elected governments. It is due to the fact that the electorate tends to evaluate the economic performance of the parties in power according to the actual state of the economy at the election time. Due basically to the division of labor, actual overall economic performance and the current rate of inflation are closely linked. So, the parties in power pursue an economic policy with an inflationary bias. By pursuing the primary goal of price stability, an independent central bank takes account of the welfare time lags involved in monetary policy. Only thereby can it provide the overall benefits which derive from price stability. It is then again a case

of paternalism so as to provide greater long run benefits at the expense of present short run benefits.

## E. Conclusion

Through both theoretical and empirical lenses, we examined some aspects of the relationship between democracy and paternalism. Weizsäcker (2024) establishes a normative foundation for merging welfare analysis and behavioral economics by presenting a preference system that acknowledges the influence of current consumption on future preferences. Welfare economics is only possible when preferences are adaptive, meaning improvements to welfare under current tastes do not undo themselves once tastes adapt. In this context, adaptive preferences may justify educative or even paternalistic guidance as a learning aid on the path to autonomy.

Laboratory experiments reveal the psychological mechanisms underlying such interventions. Although well-intentioned, people often overstep when interfering with others' choices, thereby limiting autonomy and decreasing welfare. Choice architects rely on their own (adapted) ideals as the best available proxy for learners' preferences. However, false consensus bias makes them overconfident in the relevance of their own ideals. Furthermore, people readily restrict the autonomy of others but typically reject being paternalized themselves.

Educational paternalism need not be a contradiction. When adaptive preferences make some guidance unavoidable, the ideals of projective educators can provide the social energy that drives progress. However, they must be guided by evidence of actual learner welfare. Blending Weizsäcker's macro theory with micro evidence from Ambuehl et al. and Isoni et al. helps understanding when paternalism promotes rather than diminishes autonomy.

Of course, there are forms of education that do not require direct behavioral control. Providing advice and expanding options to empower informed decisions that respect and extend one's autonomy is often possible. Ockenfels (2023), among others, provides examples of "behavioral market design" that guides behavior by extending rather than narrowing down choice menus.

Ultimately, successful democratic institutions must navigate this tension by incorporating limited, evidence-based paternalistic elements, such

as age restrictions on voting, bicameral legislatures, and independent central banks, that protect long-term collective welfare while preserving the fundamental principle of self-governance. Some paternalism is necessary to create and maintain conditions conducive to genuine democratic autonomy. The challenge is to identify the optimal boundaries of such interventions and developing mechanisms to ensure they remain accountable to democratic oversight.

## References

Ambuehl 2021, Sandro, B. Douglas Bernheim, Axel Ockenfels, What Motivates Paternalism? An Experimental Study. American Economic Review 111 (3): 787–830.

Cukierman, Alex 1992, Central Bank Strategy, Credibility, and Independence: Theory and Evidence, MIT Press, 496 pages.

Isoni 2025, Andrea, Axel Ockenfels, Robert Sugden, and Jiwei Zheng, The Demand and Supply of Paternalism, Work in Progress.

Mill 1859/1991, John Steward, On Liberty, published in Gray, John (ed), John Stuart Mill, On Liberty and other Essays, Chapter 1, Oxford, Oxford University Press.

Montesquieu 1748, Charles de Secondat, Baron de. *The Spirit of the Laws* (*De l'esprit des lois*). Geneva: Barillot & Fils.

Ockenfels 2023, Axel, Behavioral Market Design, Behavioral and Brain Sciences, 46, e171.

Samuelson 1938, Paul, A Note on the Pure Theory of Consumer´s Behavior, Economica, 5, 61–71.

Smith 1759, Adam. The Theory of Moral Sentiments. London: A. Millar, 1759.

Thaler 2008, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Weizsäcker 2024, Carl Christian von, Freedom and Adaptive Preferences, London, Routledge, XIV + 228 pages.