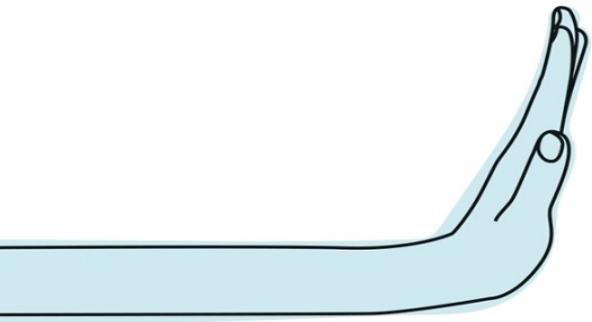
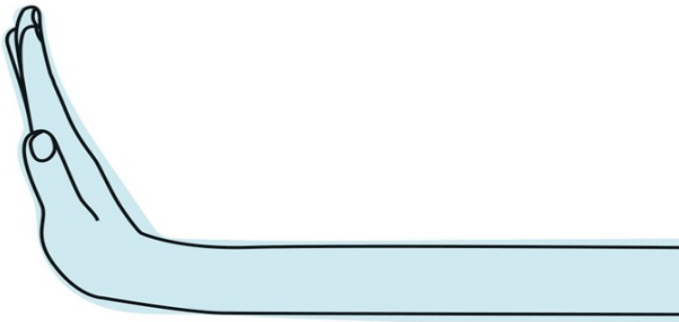
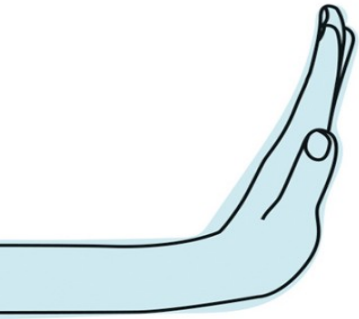
GUARDRAILS

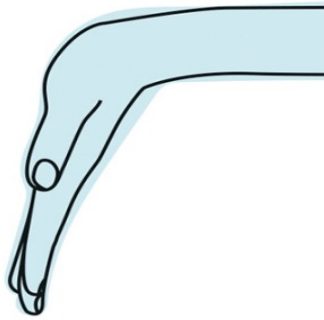


GUIDING HUMAN DECISIONS





IN THE AGE OF AI



URS GASSER &
VIKTOR MAYER-SCHÖNBERGER

GUARDRAILS

GUARDRAILS

GUIDING HUMAN DECISIONS IN THE AGE OF AI

URS GASSER

VIKTOR MAYER-SCHÖNBERGER

PRINCETON UNIVERSITY PRESS

PRINCETON & OXFORD

Published by Princeton University Press

41 William Street, Princeton, New Jersey 08540

99 Banbury Road, Oxford OX2 6JX

press.princeton.edu

ISBN 9780691150680

ISBN (e-book) 9780691256351

Version 1.0

British Library Cataloging-in-Publication Data is available

Editorial: Bridget Flannery-McCoy and Alena Chekanov

Production Editorial: Sara Lerner

Text and Jacket Design: Karl Spurzem

Production: Erin Suydam

Publicity: James Schneider and Kathryn Stevens

Copyeditor: Karen Verde

To our teachers

Herbert Burkert and Hannes Pichler

CONTENTS

GUARDRAILS

1

DECISIONS

July 1, 2002, was a dark summer night at the German/Swiss border. Well above the clouds, a Russian Tupolev 154 airliner was cruising westward. Inside it, dozens of gifted children from Ufa, southwest of the Ural Mountains, were looking forward to a holiday in Spain. In the cockpit, highly experienced captain Alexander Gross had the controls, assisted by four colleagues. Not far away, a Boeing 757 freighter was flying northward to Brussels at the same altitude.

Noticing the converging flight trajectories, an air traffic controller for Swiss air space contacted the Tupolev crew to resolve the issue. He instructed Gross to descend and the Tupolev's crew complied.

However, both airplanes were equipped with automatic collision warning systems. Just after the air traffic controller had issued his command to descend, the collision warning systems instructed both crews to take evasive maneuvers—but it ordered the freighter to descend, and the Tupolev to climb.

Having received conflicting information from the human air traffic controller and the automated collision warning system, the Tupolev crew debated whether to continue its descent or climb instead. Their discussion

was interrupted by the air traffic controller instructing them again and this time urgently to reduce its altitude, unaware that the automated system was now issuing contradictory instructions. As the crew continued on its downward trajectory—heading straight for the freighter which, following the orders of the automated system, was also descending—the warning system in the Tupolev more strongly commanded Gross to climb.

Collision warning systems in airplanes close to each other get in touch automatically and hash out which airplane is to climb and which to sink, to guarantee sufficient spatial separation between them as long as the system's commands are followed strictly. Hence, today standard operating procedures mandate that commands of the collision warning system must be complied with immediately, even if contradicting human air traffic controllers. But at the time, the pilots' training was not entirely clear on this matter. Forced to choose between human and machine, Gross chose to rely on the human controller. Shortly thereafter, at around 35,000 feet, the Tupolev collided at full speed with the Boeing freighter. Everyone on board both planes perished that night, high above the German city of Überlingen.[1]

The accident was quickly blamed on the air traffic controller, who was overworked and with some equipment not fully functional. But there is a more fundamental issue at play. On that fateful night, the Tupolev crew faced a consequential decision: Should they trust the information coming from the human controller or the collision warning system?

True, without the air traffic controller's mistaken information to descend, the crash would not have happened. But the midair collision wasn't caused only by bad information. Gross knew he had to *choose* between good and bad information, he just was unsure which was which. Rather than asking the air traffic controller for clarification or following the warning system's advice, he *chose* to descend.

Like pilots, we too face many decisions every single day, although few of them are similarly consequential. In deciding, we rely not only on information and our own thinking. Our decision-making is also shaped by

external forces, especially society, prodding, nudging, or pushing us toward a particular option, like the collision warning system. We call these *guardrails*—and that's what this book is about, from the enablers and constraints of the information we receive to rules and norms that shape how we choose among our options and how bound we are by the choices we make.

The concept of such societal guardrails is a metaphor borrowed from the kind of physical structures you see along the sides of roads or boats. Done well, these structures offer the best of both worlds. They show you where the edge is, making it less likely that you'll step over without meaning to. But they aren't like prison walls, which make it impossible to climb over if you want. You can still go off road or take a swim if you desire. Guardrails are more about marking zones of desirable behavior rather than pushing narrowly for a single "right" choice.[2]

Decisional guardrails are the interface between a person's choice and the input of society. They link the individual and the collective. Decisions taken by individuals or small groups can shape the lives of many others, as the midair crash above Überlingen so horrifically exemplifies. In a world in which decision-making is largely individual, decisional guardrails are society's most direct way to influence our mutual trajectory. This book details how, collectively, we aim to alter the decisions that are being made. It is about how society governs the contexts in which individuals make decisions—a topic both powerful and ubiquitous, yet rarely understood comprehensively.

Selecting the appropriate qualities for these decision guardrails is critical. But we will argue that in our digital age we are too quick to opt for certain types of guardrails. Without much reflection, we amplify some guardrail qualities as we overemphasize the role of technology, reflecting a widespread trend for technology to increasingly govern all kinds of human decision-making. The 2002 midair collision over Überlingen seems to confirm these beliefs: If only humans follow machines, disasters are avoided.

In this book, we suggest that such a strategy is deeply flawed. This is not because technology is somehow unable or unfit to provide effective decision governance, but because the real issue is not the nature of the decision guardrails—whether they are technical or social—but the principles underlying their design. The real question is: What kind of decisions do guardrails facilitate and what decisions should they enable?

In the nine chapters that follow we examine guardrails in a variety of challenges, contexts, and cases. But our aim is not to examine every aspect or offer a detailed blueprint; we train our eye on what we think is an emerging bigger picture—a crucial red thread in appreciating the importance of designing good guardrails. Our goal is twofold: to broaden our normative horizons, so that we realize the breadth and depths of the solution space of possible guardrails; and to offer guidance that can help us craft and select guardrails that are fitting for our challenging times—to ensure not just human agency, but human progress.

Before we can fashion a solution, however, we need to better understand what's at stake and why.

Choices, Choices Everywhere

We all make decisions—hundreds, even thousands of times every day.[3] Most of these decisions are trivial. We make them quickly and without much thinking. For others, often more consequential ones, we spend hours agonizing. Each decision shapes our future. The academic field of decision science is relatively young, having formally been established in the twentieth century. The quest to make good decisions, however, is as old as the human capacity to reflect on the choices we face.[4]

Relevant information is an obvious and crucial element of good decision-making. We glean insights from our social interactions with others, aided by the evolution of language. Script made it possible to preserve knowledge across time and space. Libraries, a cultural invention built on reading and

writing, have served for many centuries as crucial social institutions enabling us to collect information, learn from it, and use it to make life better.[5] The information stored and curated in these vast collections shaped decisions that led to important advances in areas as diverse as agriculture, architecture, medicine, art, manufacturing, and war. In the United States, libraries were assigned a crucial role at the birth of the nation: The Library of Congress was tasked with collecting the world's knowledge, and a nationwide system of public libraries aimed to bring this knowledge to the people.[6] The US Constitution makes clear that information is preserved and made available for a purpose, much as patents are granted not to reward the inventor, but "to promote the progress of science and useful arts."[7] It recognizes that the role of information, in all its mediated forms, is deeply utilitarian—improving individual and societal decisions.

More recently, digital technologies have dramatically promised to lay the groundwork for better decisions by unlocking the power of computing, data, and algorithms. More than ever before, information is at the center of our daily decision-making: We consult Siri about the weather forecast, ask ChatGPT for a couple of dinner jokes, and heed Tinder's recommendations for our next date. And indeed, in the grand scheme of things digital tools have improved the conditions for decision-making, from search engines to forecasting the spread of a virus to detecting credit card fraud from subtle anomalies in transaction data.

Information we receive needs to be analyzed and evaluated. We constantly "frame" information through our mental models about how the world works, often without much conscious thought. This is what we mean when we say that we put information into perspective. This process enables us to generate and compare options.[8] We tend to evaluate options for hugely consequential decisions more carefully, although our judgment isn't perfect —but sometimes we also fret over trivial decisions or choose bluntly without much consideration. As we ponder options, we wonder how irrevocable our

actions will be. Are we bound by them, or could we reverse course if necessary?

Pop psych literature and management training courses offer a plethora of tools and tricks to help us in this process of generating and evaluating options. We are told to "think outside the box," or make a list of pros and cons. Not every such suggestion is backed up by solid research. We can't think outside the box, for instance, in the sense that we are always thinking within mental models (and decide badly if we try without them).[9] But many suggestions may be useful in appropriate contexts.

At this point some notes of caution are in order. We are focusing here on the elements of human decision-making and how to improve that process. But we are not suggesting that all our decisions are carefully thought through. While much of our argument applies for all decision contexts, it is strongest and most valuable when we decide deliberately.

Neither are we implying that decision-making is a clean linear process, with one step followed logically after the other: collect information, analyze it using our mental models to generate decision options, compare, and choose between them. On the contrary, these elements are linked in many ways. Even deliberate decision-making is often messy and iterant. For instance, as we compare options, we may realize we missed an important dimension and must go back and gather additional information.

Nor are we suggesting that even deliberate decisions are entirely rational. Research has impressively shown that our decision-making is shaped by cognitive biases that influence our thinking. We cannot switch them off—at least not easily and at will.[10] This realization may shatter any simplistic hope that we can achieve objective rationality in the choices we make, but it isn't fatal to the idea that the decision process is open to improvement toward better reasoning.

Decisions are important because they prepare us to take actions that shape the world. But it's not just that decisions change the world—it's that *we* change the world that way. Decisions are expressions of human agency—

of our ability to influence the trajectory of our own existence and that of our species, even if only slightly. Human agency makes us matter. Without it, there would be no motivation to act. Agency is the source of energy that gets us out of bed in the morning to weather the storms of our daily lives.

Of course, we do not know whether we really have agency. Perhaps, from the vantage point of an omniscient objective bystander, both our actions and our sense of agency are just the results of biochemical processes over which we have no control.[11] But for us, the view of the nonexistent bystander is largely irrelevant. What matters, pragmatically speaking, is what we perceive every time we select an action and take it. Consequently, in this book we embrace human agency as something that we experience as existing.

Guardrails as Governance

Decisions are the cognitive mechanisms through which we interact with the world. Much hinges on them. Understandably, society has taken a keen interest in facilitating that we decide well.

Information is an important ingredient for good decision-making. And so, a variety of guardrails exist that shape what information is available. For instance, in the United States, corporate disclosure laws limit what a company's executives can share publicly and when.[12] Share too much information and you risk being fined, as Elon Musk found out when he tweeted about taking Tesla, a listed company, private in 2018.[13] In other contexts, the reverse is true, and one is required to make public certain information. Pharma companies need to disclose possible side effects for the drugs they manufacture, car companies need to publish emissions and fuel efficiency figures, and the food industry needs to put nutritional labels on most of their products.[14] Sometimes, such a *l'obligation d'information*, as the French call it poetically, may apply to a company's clients. Insurance policies are an example. The insured is typically under a duty to disclose material facts that affect the risk to the insurer. In a similar vein, the state

itself makes available a wide variety of information to help individuals make better decisions.[15] Laws are made public so that citizens can obey them, at least in democratic states. Public registers, such as for corporations or landownership, help people decide whether to engage in a business transaction.

It is not only legal rules or government policies that mandate the sharing of information. It could also be a social norm, rooted in culture and custom, such as conflict-of-interest statements in academic publications. Or it could be a practice an organization voluntarily submits to. Think, for instance, of corporate disclosure of social and environmental responsibility metrics.[16]

The hope behind all such interventions is that providing relevant information leads to better choices. When IKEA provides detailed instructions on how to assemble their furniture, they hope it will lead to decisions that make one's sofa bed more stable. When regulators mandate labels on food wrappers, they hope information about high calories and excessive amounts of sugar will lead people to make nutritious choices—though the chocolate bar might still be too hard to resist.

In the preceding examples, information is required in situations where a decision is imminent. In other contexts, information is meant to serve as a foundation for actions further down the road. It becomes an accountability tool with a longer shelf life. For instance, freedom of information mandates, so the theory goes (as usual, myriads of practical issues mess with the theory), enable citizens to make better decisions about the policies that affect their lives and, ultimately, give a thumbs up or down when the government is up for reelection.[17] Ralph Nader, the famous US government reform and consumer protection advocate, summarized it succinctly: "Information is the currency of democracy."[18]

Beyond facilitating the flow of information, guardrails extend to the process of creating and weighing decision options. For example, numerous legal rules aim to ensure that individuals can decide without undue duress, including making extortion and coercion criminal offenses.[19] In some

countries, certain particularly consequential transactions must be done before public authorities or involve testimony from experts to make sure that all parties are aware of and have considered all effects.[20] Nowhere is this more evident than in the growing number of nations that have chosen to permit assisted suicide. The decision to end one's life is so grave that these societies require multiple formal steps to confirm that the decision is deliberate, free of duress, and often in the context of a terminal and painful illness.[21]

Sometimes long-term decisions come with waiting times or "cooling-off" periods to give people ample opportunity to carefully think through their choices.[22] Being bound by a decision for a long time may have benefits—it offers stability. But we might want to think harder about whether it is the right option—and we may need more time to do so. In numerous other instances, societal guardrails explicitly enable decisions to be retracted and minds to be changed, even if that causes headaches for other parties involved.[23]

As with guardrails on information flow, guardrails on weighing options cover a spectrum from community practices to formal legal requirements. The standard operating procedures for aircraft pilots we mentioned at the start of this chapter—including whether to follow the commands of the collision warning system or the air traffic controller—are not formal law, but airlines require their flight crews to adhere to them. Similarly, emergency doctors in many hospitals must work through standard protocols of diagnosing and treating patients. It's not the law, but part of the organizational and professional culture—and it has been shown to be highly effective.

Such codes of conduct exist for many professions and organizations. Ever wondered how Amazon or McDonald's handles transaction complaints? They have detailed rules for how a customer service rep may decide and under what circumstances. Among merchants more generally, rules evolved over centuries that set out how they ought to behave when interacting with

each other. Stemming from annual trade fairs in European cities from the thirteenth century, these rules, sometimes called “lex mercatoria,” aimed to enhance trust in the market overall.[24]

A far more subtle shaping of individual decision processes has become popular lately in some policy circles. Called “nudging,” the idea is to delicately prompt people to choose the option that will be most beneficial for them. For example, when it is judged that not enough individuals opt into a retirement savings plan, one could make participation the default and require those who do not want to partake to actively opt out instead.[25] Advocates tout nudging as less limiting than more outright restrictions, but skeptics point out that nudges are opaque, creating an illusion of choice while manipulating the decision process.[26]

Similar techniques can be used to shape decisions in ways that further the interests of people other than the decision-maker. Ads and salespeople use a wide variety of cognitive tricks to influence transaction decisions.[27] Even the layout of supermarkets is carefully designed to affect our purchasing choices.[28] Deep-rooted social and cultural practices can be deliberately repurposed to shape our decisions. In the early years of eBay, sellers often rated buyers highly *before* a transaction had been completed. That didn’t make sense. Why should you rate somebody before you know whether she did as promised? Researchers took a closer look and discovered that such a premature positive rating was perceived by the buyer as a gift, which gave rise to a social expectation to reciprocate.[29] Those who quickly rated the other side in positive terms got more favorable ratings in return, which somewhat divorced ratings from the underlying transaction and prompted eBay to change its rating system.

So far, we have drawn a distinction between measures that shape the information we receive and measures that influence how we evaluate decision options. The distinction is artificial, in the sense that all measures that shape our decision processes involve information—otherwise they would not be able to reach into our mind. Airlines’ standard operating