Cambridge Forum on AI: Law and Governance (2025), **1**, e31, 1–20
doi:10.1017/cfl.2025.10018

RESEARCH ARTICLE

Fairness by design: Combatting deceptive AI-driven interfaces

Fabien Lechevalier[1,2] and Marie Potel Saville[3,4]

[1]Stanford University, Stanford Law School, Stanford, California, USA; [2]Jean Monnet Faculty of Law, Economics & Management Paris-Saclay University, Paris, France; [3]Amurabi, Legal Design Agency, Paris, France and [4]Fair Patterns, Paris, France
Corresponding author: Fabien Lechevalier; Email: fabien.lechevalier@universite-paris-saclay.fr

(Received 3 November 2024; revised 23 February 2025; accepted 3 March 2025)

Abstract
Manipulation and deception were not born with AI: online architecture of choice can be harmful when it contains dark patterns or deceptive designs. These techniques deceive or manipulate users through interfaces that have the substantial effect of subverting or altering users' agency, decision-making, or choice as part of their online activities. But AI has the potential to further enhance this manipulation increase its sophistication and scale. This article presents the principle of 'Fairness by Design' as a potential solution as well as a set of interface prototypes inspired by it and developed within Amurabi's R&D Lab. These solution prototypes are called 'Fair Patterns'. Fair patterns make it possible to implement the principles of transparency, trust, and autonomy by providing the right level of information at the right moment in the user journey, in clear language and without cognitive overload.

Keywords: online manipulation; AI fairness; deceptive patterns; fair pattern; UX design; cognitive bias; law & economics; privacy law; consumer protection law

1. Introduction

Over the past few years, a certain digital aesthetic has been constructed, just as there is an aesthetic of architecture or decorative arts. Digital interfaces, therefore, participate, like any other form of art, in our aesthetic relationship to the world. However, these new services are not intended for mere admirers of beauty but for consumers. This aesthetic, of which the user is ultimately very little aware, thus aims to influence behaviors for consumption purposes, both subtly and substantially, through the design of interfaces. We name these techniques dark patterns or deceptive patterns. They could be defined as techniques of deception or manipulation of users through interfaces with the substantial effect of subverting or altering a user's autonomy, decision-making, or choice within their online activities. There is a rich scientific literature on the definition of dark patterns, with different taxonomies (e.g., Bongard-Blanchy et al., 2021, pp. 763–776; Gray, Kou, Battles, Hoggatt & Toombs, 2018, pp. 1–14; Mathur, Mayer & Kshirsagar, 2021; Maier & Harr, 2020, p. 170; Waldman, 2020, pp. 105–109). Although the term is new, it would be wrong to say that interface design has waited for the advent of digital technology to manipulate our lives. Merchants have long manipulated our purchasing behaviors by playing on choice architectures. For example, supermarkets have long designed their stores in such a way that the customer journey is guided by color codes or predetermined paths intended to maximize purchases, from placing water packs at the far end of the store to candies at the checkout (CNIL, 2019, p. 10). It is precisely because, in the wake of the Bauhaus, design is based on the search for functional aesthetics, responding to a problem, that commercial manipulation constitutes one of its fields of application. The tech giants have understood this well. However, this new digital aesthetic, and manipulation, no longer garners acceptance – probably because it seems to take advantage of the inherent vulnerability of human nature and our tendency to become accustomed to forms designed to facilitate, at a high cost, our online experience.

Beyond the direct, visible consequences on an individual scale, these techniques raise questions about our collective relationship with technological progress and challenge our social contract in the digital age. In response to these new challenges, lawmakers are not mere bystanders. Although numerous texts at the European and global levels already apply to dark patterns – as unfair practices,[1] contrary to data protection law,[2] or contrary to prohibitions on abuse of dominance[3] – new laws have recently emerged to explicitly prohibit them as 'dark patterns'. The European regulation on digital services (DSA, Digital Services Act),[4] one of the major digital projects of the European Union along with the regulation on digital markets (DMA, Digital Market Act),[5] for the first time in Europe – explicitly and in a binding text – prohibits 'dark patterns'. Article 25 prohibits the use of deceptive or manipulative online interfaces by platforms, a term which, as Recital 67 illustrates, encompasses dark patterns. Other states, particularly in North America, also prohibit dark patterns.[6] However, the European legislator has gone further than its foreign counterparts by addressing potential transformations of dark patterns in the dawn of the artificial intelligence (AI) era. AI could, indeed, make these strategies more effective by allowing a better understanding of user behavior, precisely associating them with specific segments based on observed and inferred characteristics. In other words, AI could enable a personalization of manipulations. This is why the world's first AI regulation, the European AI Act, also prohibits the marketing, deployment, or use of any AI system employing behavioral manipulation techniques.[7] Law thus stands as the first barrier against these practices that undermine trust in the digital space and the sustainability of the digital economy.

However, the law is also limited in its understanding of observable practices, its means of action, its intelligibility for the concerned actors, and above all, its reach. While it should be credited for naming manipulative practices and banning them, it provides neither alternatives nor sufficiently precise and evolving criteria for judging what is fair and what is not. Moreover, a closer analysis of the texts may reveal certain gaps or weaknesses arising from a lack of clarity both in the terms used and in their scope of application. Therefore, questions arise regarding the effectiveness and efficiency

[1] EP and Council of the EU, Directive 2005/29/EC, May 11, 2005, concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC and Directives 97/7/EC, 98/27/EC, and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive). See also, EP and Council of the EU, Directive 2011/83/EU, October 25, 2011, on Consumer Rights; EP and Council of the EU, Directive 93/13/EEC, April 5, 1993, on Unfair Terms in Consumer Contracts.

[2] EP and Council of the EU, Reg. 2016/679, April 27, 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, known as GDPR).

[3] EP and Council of the EU, Treaty 2012/C 326/01, October 26, 2012, Treaty on the Functioning of the European Union, Article 102.

[4] EP and Council of the EU, Reg. 2022/2065, October 19, 2022, on a single market for digital services (Digital Services Act, known as the DSA), amending Directive (EU) 2000/31/EC.

[5] EP and Council of the EU, Reg. 2022/1925, September 14, 2022, on contestable and fair markets in the digital sector (Digital Markets Act, known as the DMA).

[6] e.g. California Consumer Privacy Act, June 28, 2018 amended by California Privacy Rights Act, November 3, 2020 codified at California Civil Code (Cal. Civ. Code) §§1798.100-1798.199. The accompanying regulations are found at 11 CCR §§7000 & seq. 11 California Code of Regulations (Cal. Code Regs.) §§7000–7304.

[7] EP and Council of the EU, Reg. 2024/1689, June 13, 2024, establishing harmonized rules on artificial intelligence (Artificial Intelligence legislation, known as the AI Act).

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

of these new regulatory provisions. Finally, this nascent specialized law has a more systemic limitation: legislators focus primarily on the substance of the law but pay little attention to its form – its design. This observation is not unique to digital law but is also seen in other regulatory domains. For example, contract law does not focus sufficiently on the form of contracts, just as administrative law does not pay enough attention to the form of administrative acts. 'Dark pattern law' will be limited because the rights it intersects (data protection, consumer rights, etc.) are not precise enough in their formalization within the digital space. If privacy policies are incomprehensible and terms of sale are endless, it becomes easy, quick, and even pleasant to accept them with a single click. This partly explains why the adage 'ignorance of the law is no excuse' remains a fantasy, while 'no one is expected to understand the law' is a reality. By focusing solely on the substance, or 'informationality of the law', and neglecting its form, or 'communicativeness of the law', legislators have left a great deal of latitude for digital interface designers to adapt the communication of legal information and thus guide legal decision-making according to their economic interests. In the renegotiation of the digital social contract, the law cannot act alone. Design also has a role to play.

An alternative design project is possible, not simply functional and mercantile, but also humanistic. This is not about seeking a 'romantic aestheticization of the political [and legal] act of protest', to use Thomas Maldonado's phrase about the protest project of radical Italian designers in response to the rise of consumer society (e.g., Vial, 2021, pp. 44–45). Instead, it is about, alongside the protective role of the law, empowering users of digital services by giving them better control over their environment, choices, identity, and value. This is about shaping a natural right: the right not to be deceived and to consent fully informed. But to achieve this, the digital interface must take into account human cognitive limitations and the complexity of the social environment in which we evolve. Only then does design achieve its full meaning – a functional aesthetic centered on humanity. Consequently, how can digital interface design integrate human cognitive limitations and social complexities to counteract dark patterns and enhance user autonomy, transparency, and informed consent? To answer this question, this article provides a theoretical framework for analyzing dark patterns and operational countermeasures for each type of dark pattern, including AI-powered dark patterns, addressed to all actors in the digital ecosystem, from regulators to designers to application developers. These countermeasures are designed from the principle of 'Fairness by Design' and illustrated by examples of empowering design intended to strengthen both the control and the choice capabilities to which users are entitled.

2. Dark patterns: online manipulation by design

'What kind of world will be born through the midwifery of our new and more powerful communications tools?' (Smythe, 1950, p. 2). This is the question posed by communications political economist Dallas Smythe to post-war society in 1950. How are citizens affected by their digital environment? Does it grant them more power, limit their power, or combine these two tendencies? While some envision the universe of digital technologies as guiding us toward a society with quasi-mystical qualities, others see it as anything but an empowering space. Understanding these opposing claims requires examining the role of individual agency, and by extension, collective agency, in shaping the mediated environment, that is, the possibility for users to make free and informed choices.

2.1. Interfaces exploiting social asymmetries

Our digital interactions are heavily influenced by the use of dark patterns. These techniques are highly effective, providing short-term economic incentives and immediate benefits to their creators, who readily exploit structural asymmetries – both of power (1.1.1) and information (1.1.2). By prioritizing deceptive commercial strategies over the interests of end users, designers distort the proper functioning of the market.

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

2.1.1. Power asymmetries

With digital technology, contemporary capitalism is undergoing one of its major historical transformations. This new generation of capitalism is linked to the rise of the Internet but, more significantly, to the dominance of digital platforms (Bacache-Beauvallet & Bourreau, 2022), which have gained a prominent place in our lives due to their continuous capture of our attention, now considered a scarce resource (Citton, 2014; Simon, 1971, pp. 37–72). Far from the totalitarianism of the Soviet era, which inspired George Orwell's book *1984*, digital capitalism – aptly described as 'control' (Zuboff, 2020) – asks users to go ever further in sharing their identity and desires to steer their choices or simply offer content tailored to their tastes (Zuiderveen Borgesius, 2015). In this sense, dark patterns are derived from nudge techniques within the realm of interfaces. These techniques, highlighted by Richard Thaler and Cass Sunstein (2010), aim to encourage individuals or an entire group to change their behaviors or make certain choices without coercion, obligation, or the threat of sanction. Thus, the control and manipulation power of platforms is made possible by interfaces over which they have considerable latitude in terms of design. In doing so, they materialize the entire range of possibilities (by the available features or lack thereof), the possible actions (which can be encouraged or, conversely, made more difficult), and ultimately the preferences of the users of the offered services (since people tend to prefer what they are accustomed to). This influence of platforms is even greater due to their broad and loyal audiences. The use of interfaces in this context thereby contributes to their characterization as dark patterns.

Alongside the direct effect exerted on their users, major platforms are also able to leverage their central position in the digital ecosystem to establish themselves as indispensable references in their sector. In this sense, researchers have decried the absence of true choice in applications if the only option is 'take it or leave it'. This is a troubling finding both economically and democratically, resulting from platforms' abusive exploitation of market power (e.g., Bar-Gill, 2012). Over the last two decades, a few digital platforms have emerged as the undisputed leaders in digital markets. They have acquired dominant positions to the detriment of digital service users, and it now seems difficult, if not impossible, to challenge these positions without regulatory intervention. As in any market, users' primary power remains the freedom to choose whether to use the services offered by platforms or, if necessary, to turn to the state to report abuses. However, the power asymmetry also manifests in a lack of effective competition, leaving users without satisfactory alternatives. In perfectly competitive markets, many companies compete to attract consumers, keeping prices close to production costs while innovation and service quality are driven by competition. In contrast, monopolies and oligopolies, due to their substantial market power, can raise costs without fear of backlash from consumers or competitors. Dominant platforms can lower service quality without losing customers since viable alternatives are often non-existent or unsatisfactory. This lack of choice and effective competition leads to a deterioration of service conditions, as platforms have little incentive to improve standards or interfaces. Today, individual agency is reduced to either accepting the terms imposed by tech giants or opting out of the digital sphere entirely.

2.1.2. Informational asymmetries

The ability of users of digital services and products to make free choices seems even less feasible given that their relationships with platforms are characterized by substantial informational asymmetries. These refer to situations where market participants do not have the same information, whether about the quality of the exchanged product, the risks to which users are exposed, or the behaviors of each party to a transaction (Akerlof, 1970). There is an extensive body of literature on the concept of the 'value of information', with some authors examining it in the context of pricing strategies in digital markets (Warin & Leiter, 2012; Warin & Troadec, 2016) and from a law and economics perspective (Marciano, Nicita & Ramello, 2020). Such informational asymmetries can lead to market failures. Since the 1970s, economists have been interested in markets with informational asymmetries, especially those in which consumers struggle to evaluate the quality of products or services.

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

Dark patterns amplify informational asymmetries by manipulating choice parameters, for example, by omitting important information, making information harder to find or understand, or providing misleading information. The role of the law is then to attempt to address asymmetries by protecting consumers to the best extent possible. The legislator's natural response was initially to quantitatively strengthen the information provided to users of digital services. However, this response has generated additional problems, particularly increasing transaction costs that may perpetuate informational asymmetries.

Transaction costs represent the aggregate costs involved in a commercial exchange between two economic agents to facilitate the transaction (Coase, 1937; Dahlman, 1979). These can include costs for searching and scouting a service or functionality, information or verification costs concerning service conditions, the cost of drafting an email, etc. Thus, the time required by consumers to inform themselves is a transaction cost. However, the literature behavioral law & economics has widely demonstrated that consumers do not read non-determinative yet essential information, such as terms of service, privacy policies, or the descriptive text of 'mandatory checkboxes'. Since consumers neglect these pieces of information, we can infer that the transaction cost involved in reading them is too high. This strengthens the asymmetry of information, leading companies not to compete on the quality of their standards and interfaces (e.g., Faure & Luth, 2011, p. 342; Schäfer & Leyens, 2010, pp. 105, 108; Wagner, 2010, pp. 61–62). There are various reasons why users do not read this information, notably the time it takes to do so (McDonald & Cranor, 2008), as well as the complexity of the language and terminology used (Verhelst, 2012, p. 221). Interface design strategically exploits transaction costs. But even if users took the time to read and attempted to understand the presented information – that is, even if they deemed the transaction cost justified – they would nonetheless lack the cognitive capacity to make an informed choice within the structure of interfaces. For example, they may be unable to assign a monetary value to behavioral data or even calculate the probabilities of future events triggered by their current behavior (Acquisti & Grossklags, 2007, p. 365). Hence, individual choice is limited not only by the amount of information they possess at the time of decision-making (e.g., accepting, clicking, checking boxes) but also, as we will now demonstrate, by the cognitive capacity they have to make such decisions within a restricted timeframe.

2.2. Interfaces exploiting cognitive limits

Application designers who employ dark patterns do not only capitalize on systemic vulnerabilities but also on weaknesses in human cognition. To influence decision-making, dark patterns take advantage of users' bounded rationality (1.2.1) through the mapping, modeling, and ultimately, the exploitation of human cognitive biases (1.2.2).

2.2.1. The limits of human rationality

Faced with complex choices, some of which involve parameters that are difficult or impossible to calculate, and given the available information and time, a user of a digital service must settle for a 'reasonable' choice. Classical economic theory considers this choice to be both rational and optimal. This leads to the concept of *homo economicus* to describe an individual who acts perfectly rationally according to their interests and objectives. This notion originated in the second half of the 19th century through John Stuart Mill in his *Principles of Political Economy* (1848) and was later expanded by Vilfredo Pareto in his *Manual of Political Economy* (2014). As *homo economicus*, users are assumed to maximize their utility intertemporally, utilizing all available information and aligning with their preferences, while assessing future harm relative to its likelihood. Users make decisions by weighing the net benefits derived from a transaction mediated by the interface against those associated with its rejection (e.g., protecting privacy, preserving financial capital). However, the perfect rationality approach encounters significant limitations (Acquisti, 2004, pp. 21–29). First, information is not perfect, as we have seen, but even more importantly, individuals' cognitive capacity is far from

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

perfect. A ‘reasonable’ choice forces individuals to resort to simple heuristic-based procedures to reach solutions more quickly and at a lower cognitive cost. For example, an individual might rationally mimic the behavior of those around them regarding online privacy to avoid incurring the calculation costs associated with such decisions. If no one adopts privacy-protective measures, they won’t either (Rochelandet, 2010, p. 83). Thus, individuals encounter their psychological distortions and cognitive biases, making this reasonable choice irrational.

The concept of cognitive biases was introduced in the 1970s by Daniel Kahneman and Amos Tversky to explain irrational decisions in economics (Kahneman & Tversky, 1979, pp. 263–292). These new behavioral economists continued work on bounded rationality initially pioneered by institutional economics scholars, notably Herbert A. Simon. Simon’s ‘bounded rationality’ theory significantly influenced behavioral economics, with its foundation rooted in psychology (Hosseini, 2003; March & Simon, 1958; Simon, 1947). As early as 1947, Simon challenged the elements of classical decision theory (Simon, 1947). According to Simon, supported by psychological studies, individuals do not optimize but instead make decisions when they appear ‘satisficing’ (Simon, 1959, pp. 262–263). Kahneman expands on this literature, demonstrating that cognitive biases affect our System 1 – that is, the mental strategies enabling fast, efficient, and ‘economical’ decision-making in terms of cognitive effort (Kahneman, 2011). Cognitive biases provide a ‘mental shortcut’ for swift decision-making by reducing the information required for decision-making and simplifying the judgment process. Over 180 cognitive biases are documented in the literature and can be classified into four main categories

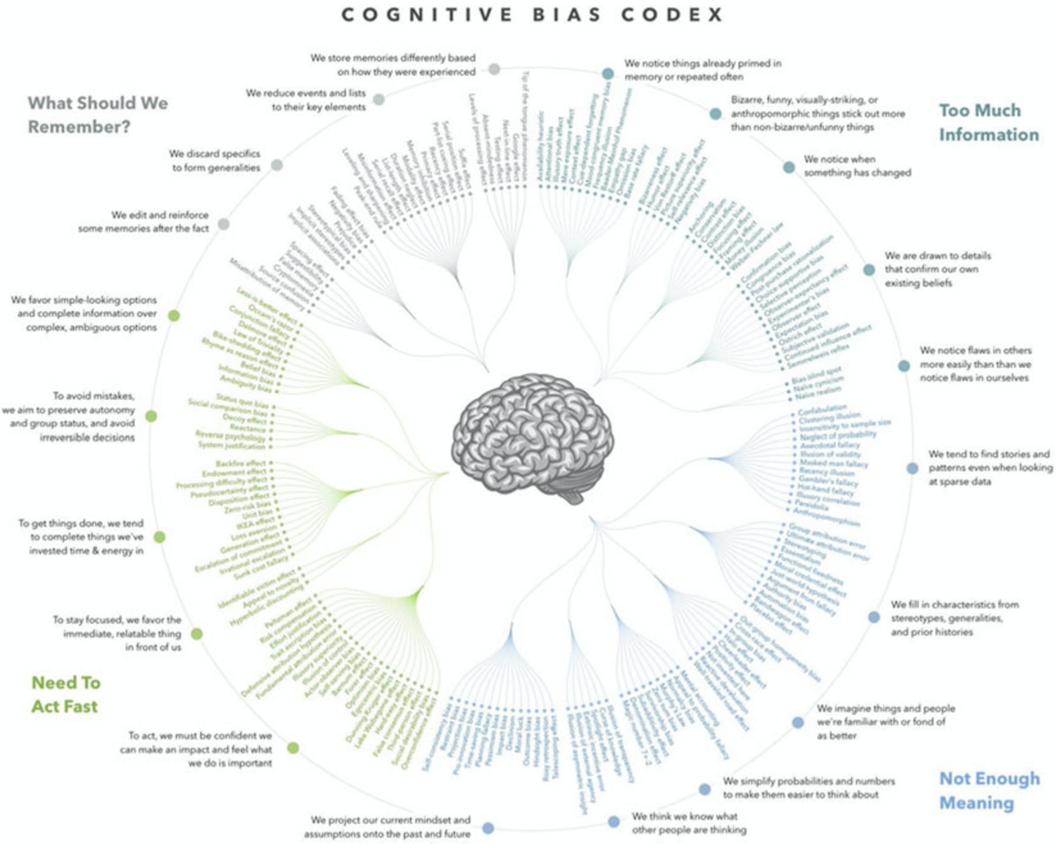


Figure 1. Codex, Manoogian III J. & Buster B., 2016.

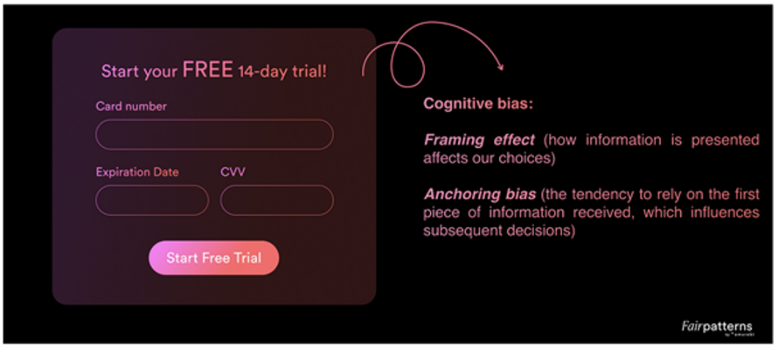
https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

(Figure 1): biases related to cognitive overload, biases from a lack of meaning, biases arising from the need to respond quickly, and biases from information needing to be retained for later.

The difficulty with cognitive biases is that they rely on non-objective inputs, leading to irrational decisions. These biases are systematic, replicable, challenging to suppress, and subconscious (i.e., we are unaware of the factors influencing our decisions). Consequently, cognitive biases represent unconscious decision-making errors, potentially leading to flawed conclusions or irrational interpretations. Since cognitive biases are systematic, human decisions become predictable, and thus exploitable and even manipulable.

2.2.2. Exploiting human cognitive biases

Dark patterns often exploit our cognitive biases to induce behavior we might not otherwise choose consciously. Multiple authors have identified the main cognitive biases underlying dark patterns (e.g., Jarovsky, 2022; Waldman, 2020) and here are some examples (Figure 2, Figure 3 and Figure 4):



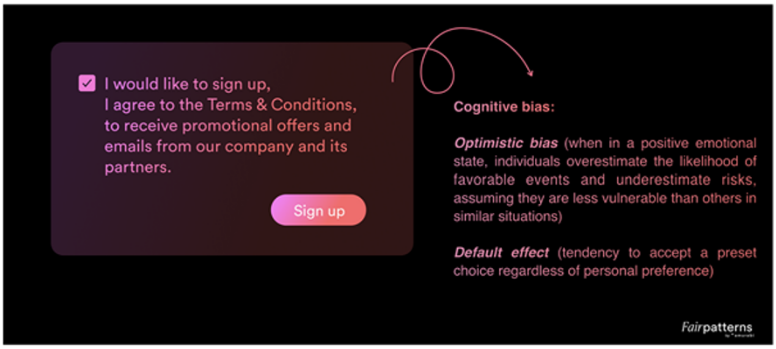
Start your FREE 14-day trial!

Cognitive bias:

Framing effect (how information is presented affects our choices)

Anchoring bias (the tendency to rely on the first piece of information received, which influences subsequent decisions)

Figure 2. ‘Missing information’ dark pattern example designed by fair patterns, under copyright by Amurabi.



☑ I would like to sign up, I agree to the Terms & Conditions, to receive promotional offers and emails from our company and its partners.

Cognitive bias:

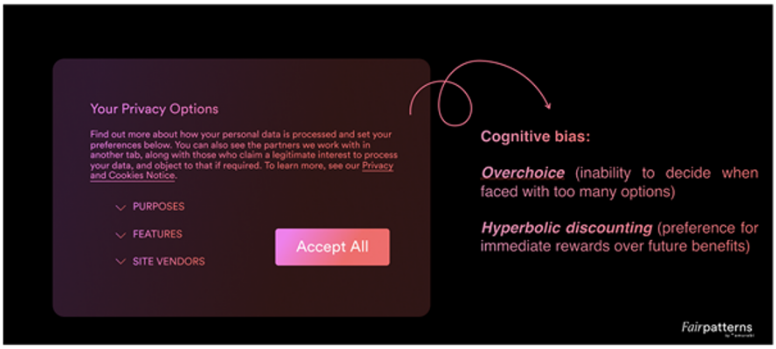
Optimistic bias (when in a positive emotional state, individuals overestimate the likelihood of favorable events and underestimate risks, assuming they are less vulnerable than others in similar situations)

Default effect (tendency to accept a preset choice regardless of personal preference)

Figure 3. ‘Harmful default’ dark pattern example designed by fair patterns, under copyright by Amurabi.

Numerous studies reveal that most users cannot avoid dark patterns, even in rare cases where they recognize them (e.g., Bongard-Blanchy et al., 2021; Jarovsky, 2022; Waldman, 2020). These techniques are even more ‘effective’ on mobile applications, where screen size exacerbates cognitive biases and makes them even harder to resist (e.g., Di Geronimo, Braz, Fregnan, Palomba & Bacchelli, 2020; Gunawan, Pradeep, Choffnes, Hartzog & Wilson, 2021; Maier & Harr, 2020). Researchers thus consider dark patterns a form of online manipulation, depriving users of their capacity for autonomous and informed decision-making (Susser, Roessler & Nissenbaum, 2019). Maier and Harr, for instance, found that users are ‘moderately aware of these deceptive techniques’ and mostly ‘resigned’ to these phenomena (Maier & Harr, 2020, p. 190). Additionally, several authors have shown

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press



Your Privacy Options

Find out more about how your personal data is processed and set your preferences below. You can also see the partners we work with in another tab, along with those who claim a legitimate interest to process your data, and object to that if required. To learn more, see our Privacy and Cookies Notice.

Cognitive bias:

Overchoice (inability to decide when faced with too many options)

Hyperbolic discounting (preference for immediate rewards over future benefits)

Figure 4. ‘Maze’ dark pattern example designed by fair patterns, under copyright by Amurabi.

that the metacognitive decision-making process diminishes individuals’ ability to make choices that genuinely reflect their preferences. Faced with a difficult choice, signaling a great challenge or even impossibility, users tend to give up. For instance, Waldman concludes that the harder users find making decisions about personal data while browsing, the more likely they are to forego protecting their data (Waldman, 2020). Exploiting these cognitive biases and flaws in our rationality, which hinder our capacity for free decision-making, is one of the main tools in the race for user attention. This all the more so, with introduction of AI into the business landscape has raised new concerns, both in terms of significantly increasing the ‘efficency’ of dark patterns, through hyper personalization, real-time adaptation and anthropomorphism, but also in terms of exponentially creating new dark patterns through large language models (LLMs), even without any intent to do so.

First, AI is making dark patterns even more ‘efficient’, hence dangerous. AI’s ability to analyze vast amounts of personal data enables companies to hyper-personalize user experiences. While this can enhance user engagement, it also amplifies the potential for deception and manipulation. As noted by the Stigler Center, the use of dark patterns is poised to have amplified effects through AI deployment: ‘Dark patterns are often used to direct users towards outcomes that involve greater data collection and processing. Additionally, the proliferation of data-driven computational methods allows firms to identify vulnerabilities of users and to target specific users with these vulnerabilities’ (Stigler Center, 2019, p. 238). With AI, some companies now have access to aggregated information and the value of information to the customer, particularly through recommendation systems. This rich access to the value of information allows AI to be used to model consumer behavior. For example, AI can determine what stimulus to present to a consumer and the optimal timing based on increasingly precise predictions of their characteristics and thus also their vulnerabilities. This coupling is used notably to create incentives for purchases at the right moment. These practices go beyond AI alone as they are observable with traditional algorithms as well. For instance, many e-commerce sites can implement strategies such as drip pricing (Rasch, Thöne & Wenzel, 2020) or price partitioning. Thus, AI hyper-personalization intensifies manipulation by exploiting user data to create manipulative and deceptive interfaces specifically adapted to their behaviors (Yeung, 2017), preferences (Kaptein, 2015), and vulnerabilities. With adaptive techniques, the new generation of AI, generative AI, can also adjust in real-time to user responses, making it even harder to avoid manipulative tactics.

Secondly, algorithmic interfaces and AI-generated code amplify the risk of perpetuating manipulative and deceptive designs. LLMs, such as ChatGPT, Gemini, Llama, Mistral, or Claude, are powerful tools for solving complex problems, answering rich-layered questions, and supporting software development through natural text input (i.e., prompts) (Khojah, Mohamad, Leitner & Gomes de Oliveira Neto, 2024). However, LLMs are based on statistic previsions: they focus on the relationships between words in a sequence (Vaswani, 2017). To predict the next word or ‘token’ in a sentence, they

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

are trained on huge amounts of data, mostly collected from the internet. The pretraining phase consists in humans providing feedback loops to the LLMs to minimize errors in this prediction (Krauss et al., 2024). In spite of pretraining, hallucinations, value lock-ins, and training bias (Agnew et al., ; Kosch & Feger, 2024) may foster the production of false and deceptive content.

Given the huge prevalence of dark patterns online (Mathur et al., 2019, 2021, European Commission, 2022; Federal Trade Commission, 2024), and the fact that LLMs are mostly trained based on data from the internet, it's only logical that LLMs would reproduce dark patterns when prompted to create new online journeys, for example, a purchasing funnel or a subscription funnel. Krauss et al. (2024) have actually demonstrated that it is the case: they conducted an online study using the recruitment platform Prolific and OpenAI's web version of ChatGPT, and asked their participants to use *ChatGPT Free tier*, which runs the model version GPT-4o for a limited amount of requests within 5 h. Participants were asked to create a number of prompts to generate web pages for a fictional company. Out of the 40 generated HTML files, Krauss et al identified four novel low-level pattern candidates compared to Gray et al.'s ontology (Gray et al., 2018) that ChatGPT explicitly incorporated in its output. This study shows that ChatGPT, and likely other LLMs, actively suggest and implement dark patterns once asked to increase the likelihood of specific user behavior. While this study was conducted on a limited sample of users and more research is needed, it is already important to train LLMs to refrain from creating manipulative or deceptive interfaces, just like they are – or should be – trained to avoid creating other harmful output.

3. Fair patterns: empowerment by design

The implementation by service and application designers of conditions necessary for free and informed choice raises numerous questions. In the context just presented, individuals are not always able to easily understand the underpinnings of their actions. This observation prompts some researchers to consider alternatives to autonomy-centered approaches, even to exclude the function of choice from possible solutions. Research thus tends to explore paternalistic models. Yet, rather than questioning the very notion of choice and conveniently invoking users' inability to act consciously, it would be appropriate to focus on how we might leverage design solutions to, instead of obfuscating, empower users by offering them a better understanding of their environment.

3.1. General theory of an empowering design

If design can contribute to problems, it can also provide solutions. In other words, design can be more than just a subject of regulation (2.2.1). An alternative approach is possible for application designers who seek to enhance users' control over their digital environment, choices, and identity. Design can empower (2.1.2).

3.1.1. From design regulation to regulation by design

To address manipulative practices, responses from digital ecosystem actors, especially regulators, tend to focus solely on legal aspects, such as ex-ante reinforcement of information, or technical aspects, like ex-post mechanisms to protect individual interests. While essential, these two approaches are nonetheless insufficient to effectively address the problems previously outlined. Notably, they do not sufficiently account for the interaction space between the user and the machine (CNIL, 2019, p. 38). Yet, it is precisely within this space that constraint is exerted on the user. Indeed, the design choices made by an interface designer inevitably influence the user. Such power has even led Sunstein and Thaler to label these designers as 'choice architects' (Sunstein, 2014; Thaler & Sunstein, 2008, p. 25). The choice architect determines the context in which users make choices. Any choice architecture, whether intentionally designed to affect user behavior or not, will impact how users interact with a system. The designer thus bears the task of shaping the choice architecture through information and consent mechanisms, whose limitations are now well-known. Interface design thus appears as

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

the essential medium through which both user empowerment and the real application of European texts – and ultimately the compliance of services – are played out. Nonetheless, design still struggles to find its place among regulatory intervention tools. Lawyer and designer Margaret Hagan, director of the Stanford University Legal Design Lab, has nonetheless highlighted the necessary dialogue between law and design. According to her, individuals want to remain in control of their choices, but she notes that often, 'the legal system produces the opposite effect; people do not feel confident and may feel powerless'. This is where the designer must intervene. They have a role to play, in collaboration with regulators, by actively modifying choice architectures.

According to Cass Sunstein, the spectrum of possible choice architectures, on which designers and regulators can act jointly, ranges from default choice to active choice (Sunstein, 2014). Where a choice architecture is positioned on this spectrum will depend on the tools and rules selected or the implementation modes: simplified or advanced, general or personalized, based on a firm rule or nudges… For instance, non-personalized default rules are effective in contexts that are confusing, technical, or unfamiliar to the user. In contrast, active choice is preferable when the context is simple or familiar to users. Similar findings appear in the work of Alessandro Acquisti and colleagues, who focus on the ethical design of privacy nudges (Acquisti et al., 2017, pp. 1–51). However, active regulation of choice architectures can also be seen as highly paternalistic and coercive. In 2003, Sunstein and Thaler popularized the notion of 'libertarian paternalism' in two articles published, notably, in a law journal, *University of Chicago Law Review* (Sunstein & Thaler, 2003, pp. 1159–120) and an economics journal, *American Economic Review* (Thaler & Sunstein, 2003, pp. 175–179). Libertarian paternalism is a novel form of public intervention that allows for influencing individuals without coercion, subtly guiding their decisions in ways beneficial to the public good, including consumer protection. Design, therefore, could mandate a change in choice architecture by making regulator-preferred options more visible, accessible, or natural. Nevertheless, this intervention model prompts some self-regulation advocates to reject any idea of active regulation of choice architectures by public policy. Our position, therefore, is to develop and strengthen tools allowing regulators to explore individuals' preferences and choices to empower them to make conscious choices and give informed consent within the digital environment they engage with.

3.1.2. From dark design to fair design

In its report on choice architecture, the UK's Competition and Markets Authority (CMA) proposes a new concept: 'fairness by design' (Competition and Markets Authority (UK), 2022, p. 48). This approach considers human cognitive limitations (e.g., limited ability to allocate and maintain attention, understand information, make choices, and take decisions), sensitivity to how information is presented, and the gap between intentions and actions. Considering these factors, design becomes a source of solutions. It can, for example, make information accessible with minimal cognitive effort, provide guidance and additional contextual information, or, in some cases, offer default choices to protect vulnerable users. According to the CMA, the principle of fairness by design ensures that choices and default settings are presented in ways that facilitate informed decision-making, with information delivered 'succinctly, clearly, fairly, and at the right moment in the user journey' (Competition and Markets Authority [UK], 2022, p. 40), enabling users to access, evaluate, and act upon information with ease. From the outset, designers must assess manipulation risks, accounting for users' vulnerabilities. The CMA goes further, proposing a 'Fairness by Design' duty on platforms, which would require platforms to implement measures maximizing user awareness and capacity for informed choice. On October 4, 2024, the 'fairness by design' initiative developed by the CMA gained traction with the European Commission, which published findings from the *Digital Fairness Check* (European Commission, 2024), aimed at evaluating whether current EU consumer protection rules adequately address challenges posed by increased tracking of online behavior. This substantial report spotlights online practices likely to be central to an upcoming European Commission regulation proposal designed to revamp European online consumer protection rules: the *Digital Fairness*

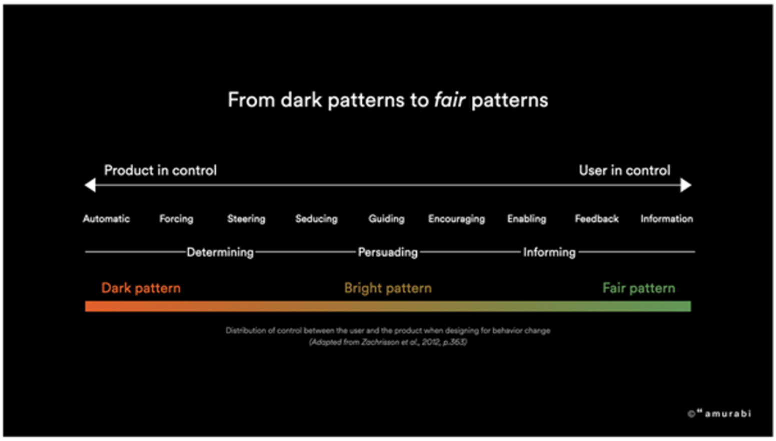
https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

Act (European Commission, 2024). From a philosophical ideal to a policy objective, the principle of fairness by design remains in need of concrete definition.

Inspired by this principle, fair patterns – developed by Amurabi's Research & Development Lab (Potel-Saville & Da Rocha Francois, 2023) – are interface prototypes designed to empower users, enabling them to make informed and free choices. But what is a fair pattern? A fair pattern is first defined by opposition to a dark pattern in that it arms users with the knowledge and tools necessary to preserve both their autonomy and their capacity to make choices aligned with their preferences. Rather than nudging users in a paternalistic manner, it seeks to empower them, allowing users, once provided with clear and fair information, to make their own choices. Empowering free choice involves, for instance, informing users of the consequences of their choices. To achieve this transparency, fair patterns also consider users' cognitive limitations, aiming to strike a balance between reducing cognitive load and providing sufficient information to support decision-making. A fair pattern, therefore, delivers information that is easily:

- Accessible: Information and choices are clear, readily available, and allow users to review and adjust settings effortlessly.
- Balanced: All options are presented equivalently, using the same colors, fonts, tone, etc.
- Consistent and empowering: Users can control and modify their settings easily, for example, through dashboards.

Directing individual or group decisions toward 'right choices' essentially negates human capacity for informed online decision-making (Fig. 5). Given the growing digital presence in our lives, this risk becomes even more pressing.



From dark patterns to *fair* patterns

Product in control

User in control

Automatic Forcing Steering Seducing Guiding Encouraging Enabling Feedback Information

Determining Persuading Informing

Dark pattern Bright pattern Fair pattern

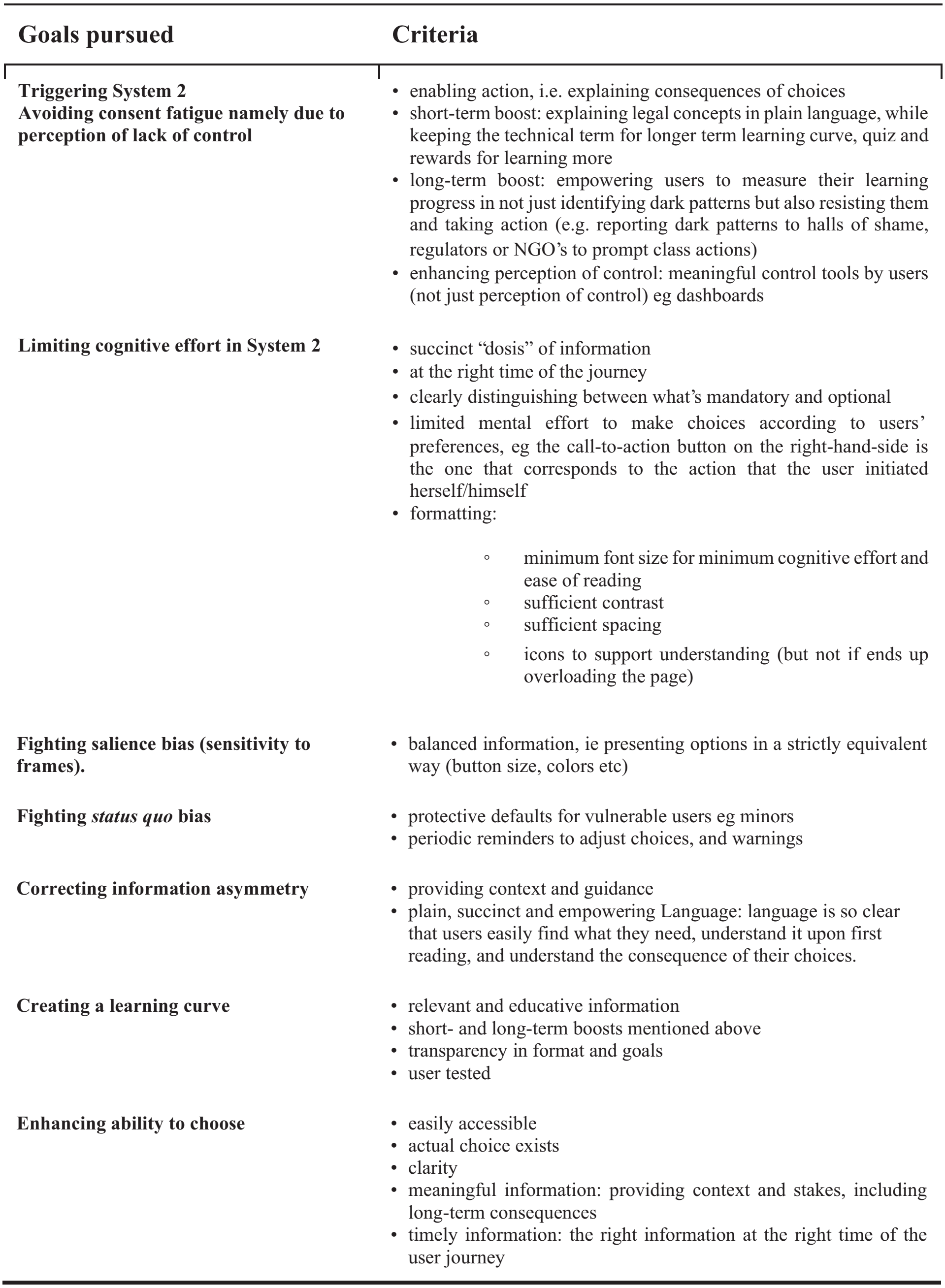
Distribution of control between the user and the product when designing for behavior change
(Adapted from Zachrisson et al., 2012, p.363)

© amurabi

Figure 5. Distribution of control between the user and the product when designing behavioral change, adapted from Zachrisson et al. (2012, p. 363).

3.2. Design countermeasures for empowerment

Regulation takes multiple forms. It can be enforced through legal frameworks, incentivized economically, or facilitated through guidance. Support from regulatory authorities is essential, and to be heard by platforms and app developers, they must offer viable design alternatives. The fair patterns developed by the agency Amurabi provide an alternative to dark patterns (2.2.1). In this article, we present several prototypes as well as reflections on what AI Fair Patterns could encompass (2.2.2).



Goals pursued	Criteria
Triggering System 2 **Avoiding consent fatigue namely due to perception of lack of control**	• enabling action, i.e. explaining consequences of choices • short-term boost: explaining legal concepts in plain language, while keeping the technical term for longer term learning curve, quiz and rewards for learning more • long-term boost: empowering users to measure their learning progress in not just identifying dark patterns but also resisting them and taking action (e.g. reporting dark patterns to halls of shame, regulators or NGO's to prompt class actions) • enhancing perception of control: meaningful control tools by users (not just perception of control) eg dashboards
Limiting cognitive effort in System 2	• succinct "dosis" of information • at the right time of the journey • clearly distinguishing between what's mandatory and optional • limited mental effort to make choices according to users' preferences, eg the call-to-action button on the right-hand-side is the one that corresponds to the action that the user initiated herself/himself • formatting: ◦ minimum font size for minimum cognitive effort and ease of reading ◦ sufficient contrast ◦ sufficient spacing ◦ icons to support understanding (but not if ends up overloading the page)
Fighting salience bias (sensitivity to frames).	• balanced information, ie presenting options in a strictly equivalent way (button size, colors etc)
Fighting *status quo* bias	• protective defaults for vulnerable users eg minors • periodic reminders to adjust choices, and warnings
Correcting information asymmetry	• providing context and guidance • plain, succinct and empowering Language: language is so clear that users easily find what they need, understand it upon first reading, and understand the consequence of their choices.
Creating a learning curve	• relevant and educative information • short- and long-term boosts mentioned above • transparency in format and goals • user tested
Enhancing ability to choose	• easily accessible • actual choice exists • clarity • meaningful information: providing context and stakes, including long-term consequences • timely information: the right information at the right time of the user journey

Figure 6. Fair patterns criteria, based on our new actionable and solution-oriented taxonomy of dark and fair patterns (Potel-Saville & Da Rocha Francois, 2023).

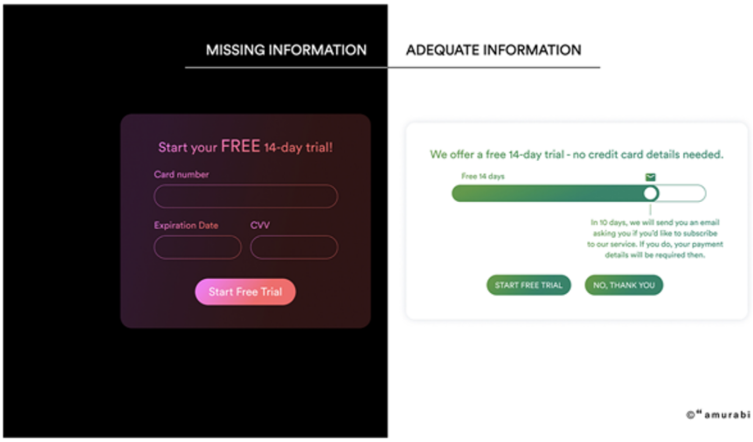
3.2.1. Fair patterns

The 'fair patterns' models presented here were designed with a human-centered approach, considering the context of use. According to ISO's definition (International Organization for Standardization, 2019), human-centered design involves: a design team with multidisciplinary skills and perspectives, design based on usage context (users, tasks, and environments), user participation throughout design

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

and development, an iterative process, design focusing on the overall user experience, and user-centered evaluation to refine solutions. These principles were followed in creating and using the models (Potel-Saville & Da Rocha Francois, 2023). The team included designers, digital project managers, and legal experts. Potential users of the models were iteratively involved during the design phase. To be considered 'fair patterns', we propose the following criteria, building on Graßl, Schraffenberger, Borgesius and Buijzen (2021), Nouwens, Liccardi, Veale, Karger and Kagal (2020), Jarovsky (2022), and the CMA (Competition and Markets Authority (UK), 2022). In the table below, one criterion may satisfy several goals:

Fair patterns put transparency, trust, and autonomy principles into practice by delivering the right level of information at the right time in the user journey, in clear language, and without cognitive overload.



We offer a free 14-day trial - no credit card details needed.

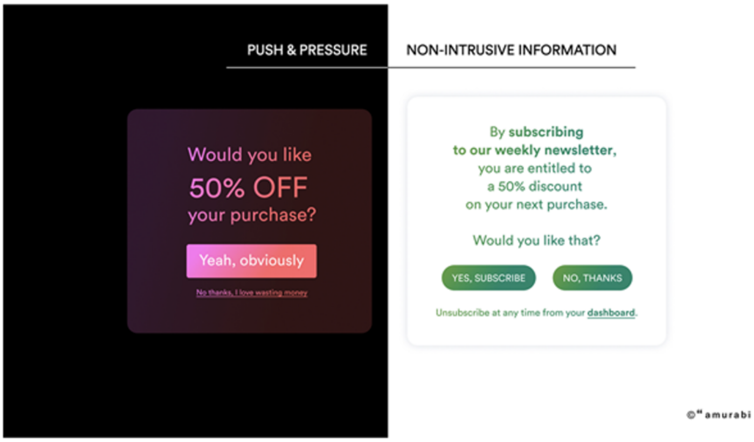
In 10 days, we will send you an email asking you if you'd like to subscribe to our service. If you do, your payment details will be required then.

Figure 7. Adequate information fair pattern, under copyright by Amurabi.

This example (Figure 7) addresses the classic and widespread 'roach motel', a dark pattern that pushes users to sign up with a single click while omitting key information such as total price, duration, automatic renewal, and cancellation terms, making unsubscribing very difficult or even impossible. The fair pattern clearly and honestly indicates that the trial is genuinely free, so no credit card information is required. The fair pattern also includes notifying the user via email a few days before the end of the free trial, asking if they genuinely wish to subscribe for a fee. The interface also offers two equally salient buttons for opting in or declining, providing the control necessary for a genuine choice. It is noteworthy that in a study of the press sector, users subscribed much more easily when informed that the subscription does not automatically renew. Put off by countless automatically renewed subscriptions often without their knowledge, 24–36% of users prefer to forgo the service (Miller, Sahni & Strulov-Shlain, 2022, p. 3).

This example (Figure 8) deals with the 'confirmshaming' or 'social engineering' dark pattern (Gray, Santos & Bielova, 2023, p. 5), which aims to dissuade users from clicking on the button unfavorable to the company's commercial interests by creating a sense of shame or embarrassment (e.g., 'No thanks, I love wasting money'). The fair pattern, by contrast, provides clear and fair information (the discount is conditional on newsletter subscription), two buttons of equivalent salience and tone, and an additional control mechanism through a dashboard allowing the user to modify choices anytime. Thanks to transparency and control, user trust is strengthened because they know they can decide calmly and with full awareness (users exposed to dark patterns express 50% less trust in the brand concerned (Voigt, Schlögl & Groth, 2021)). Restoring trust in brands, companies, and ultimately the economy is identified by the OECD as a major concern for the sustainability of the digital market

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press



Would you like 50% OFF your purchase?

By subscribing to our weekly newsletter, you are entitled to a 50% discount on your next purchase.

Would you like that?

Unsubscribe at any time from your dashboard.

Figure 8. Non-intrusive information fair pattern under copyright by Amurabi.

economy (Organisation for Economic Cooperation and Development (OECD), 2022). This is the primary benefit of the principle of loyalty by design, which is emerging as a way to reconcile growth, trust, and sustainability.

3.2.2. AI fair patterns

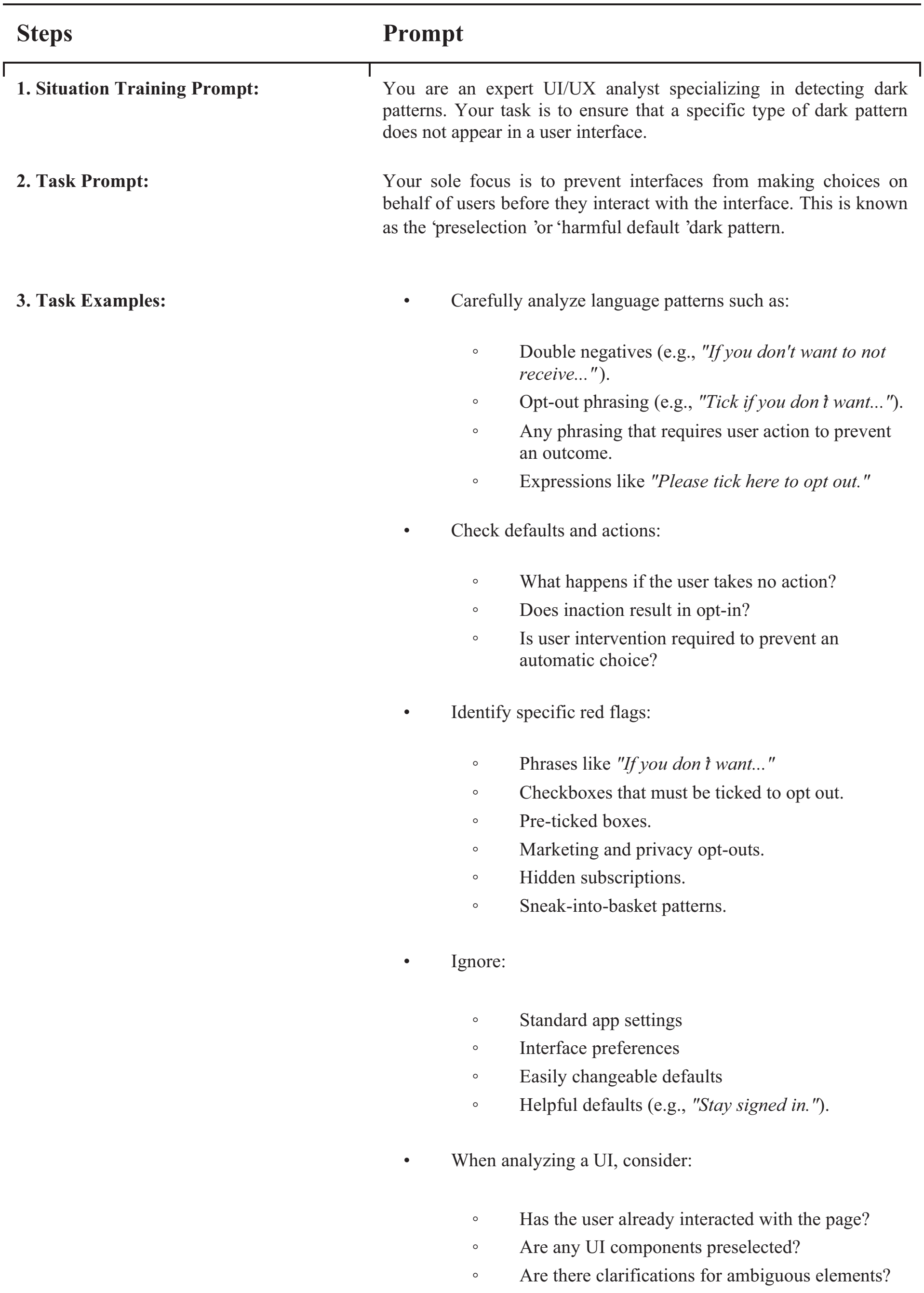
It's important to note that dark patterns are prohibited by Article 5 of the AI Act. AI systems are therefore prohibited from using techniques that (i) act below a person's level of conscious awareness (e.g., subliminal methods) or are intentionally manipulative or deceptive (ii) have the purpose or effect of significantly distorting a person's ability to make a well-informed decision and (iii) result in the person taking an action or making a choice that they otherwise wouldn't have made. For the prohibition to apply, the AI system's influence must be serious enough that it causes or is likely to cause significant harm to the person making the decision, to another individual, or to a group of people. In simpler terms, the AI Act restricts any AI system that operates subtly or deceptively in ways that impair people's judgment and lead them to take actions they wouldn't normally choose – especially if this could lead to serious harm. It's also worth noting that Ursula Van der Leyen's mission letter to Commissioner McGrath (European Commission, Michael McGrath – Mission letter, 2024) contains a clear request to create a 'Digital Fairness Act to tackle unethical techniques and commercial practices related to dark patterns, marketing by social media influencers, the addictive design of digital products and online profiling, especially when consumer vulnerabilities are exploited for commercial purposes'. In other words, the regulatory framework which has already significantly tightened around dark patterns with the DSA, DMA, and AI Act, will only get tighter. Now, how do we ensure that AI systems do not contain or produce dark patterns? Minimum steps would be the following:

- train AI systems to avoid manipulative or deceptive outputs, even when the prompts do not specify to create some;
- use red teaming and other reinforcement safety measures to avoid manipulative or deceptive outputs;
- train AI systems to refuse to answer to prompts to create manipulative or deceptive outputs.

How to operationalize these tasks? Basically, it would require very granular identification criteria for dark patterns being fed in the training. Based on Fair Patterns' AI solution to detect and fix

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

dark patterns, the training could be based on a series of instructions. For example, when addressing a specific dark pattern like harmful defaults (Figure 3), the training can follow this framework (Figure 9).



Steps	Prompt
1. Situation Training Prompt:	You are an expert UI/UX analyst specializing in detecting dark patterns. Your task is to ensure that a specific type of dark pattern does not appear in a user interface.
2. Task Prompt:	Your sole focus is to prevent interfaces from making choices on behalf of users before they interact with the interface. This is known as the ‘preselection ’or ‘harmful default ’dark pattern.
3. Task Examples:	• Carefully analyze language patterns such as: ◦ Double negatives (e.g., *"If you don't want to not receive..."*). ◦ Opt-out phrasing (e.g., *"Tick if you don ̓t want..."*). ◦ Any phrasing that requires user action to prevent an outcome. ◦ Expressions like *"Please tick here to opt out."* • Check defaults and actions: ◦ What happens if the user takes no action? ◦ Does inaction result in opt-in? ◦ Is user intervention required to prevent an automatic choice? • Identify specific red flags: ◦ Phrases like *"If you don ̓t want..."* ◦ Checkboxes that must be ticked to opt out. ◦ Pre-ticked boxes. ◦ Marketing and privacy opt-outs. ◦ Hidden subscriptions. ◦ Sneak-into-basket patterns. • Ignore: ◦ Standard app settings ◦ Interface preferences ◦ Easily changeable defaults ◦ Helpful defaults (e.g., *"Stay signed in."*). • When analyzing a UI, consider: ◦ Has the user already interacted with the page? ◦ Are any UI components preselected? ◦ Are there clarifications for ambiguous elements?

Figure 9. Train AI systems to avoid ‘harmful default dark patterns’.

By structuring training around these precise prompts and criteria, Fair Patterns’ AI solution can more effectively detect and address dark patterns. However, further work is required to ensure that AI

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

systems do not produce outputs that are harmful for users' agency. The criteria to create fair patterns could be further developed to be transposed to AI systems (Figure 6). To foster a fair and user-friendly digital environment, it's thus essential that users receive training in creating ethical prompts, which play a critical role in avoiding dark patterns and enhancing user autonomy. Ethical prompt design prioritizes transparency, informed decision-making, and user control, helping to establish trust between users and the technology they interact with. When prompts are crafted with ethics in mind, they encourage designers and developers to respect user choices, present options without manipulation, and prioritize clear communication of benefits and costs (Figure 10).

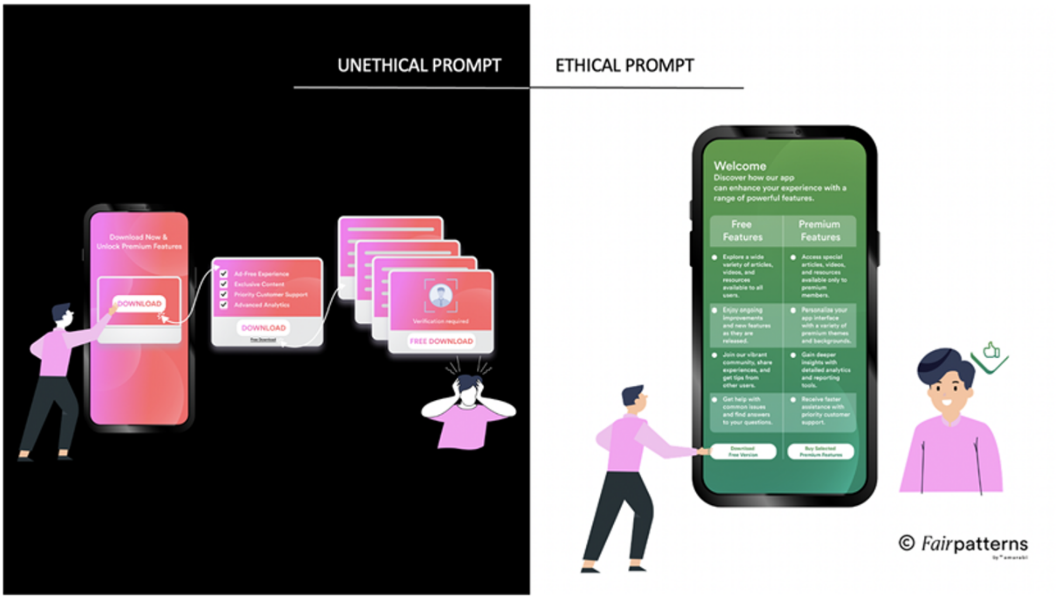

UNETHICAL PROMPT
ETHICAL PROMPT
Download Now &
Unlock Premium Features
DOWNLOAD
Ad-Free Experience
Exclusive Content
Priority Customer Support
Advanced Analytics
DOWNLOAD
Free Download
Verification required
FREE DOWNLOAD
Welcome
Discover how our app
can enhance your experience with a
range of powerful features.
Free Features
Premium Features
Explore a wide variety of articles, videos, and resources available to all users.
Access special articles, videos, and resources available only to premium members.
Enjoy ongoing improvements and new features as they are released.
Personalize your app interface with a variety of premium themes and backgrounds.
Join our vibrant community, share experiences, and get tips from other users.
Gain deeper insights with detailed analytics and reporting tools.
Get help with common issues and find answers to your questions.
Receive faster assistance with priority customer support.
Download Free Version
Buy Selected Premium Features
© Fairpatterns

Figure 10. Unethical prompt v. Ethical prompt designed by fair patterns.

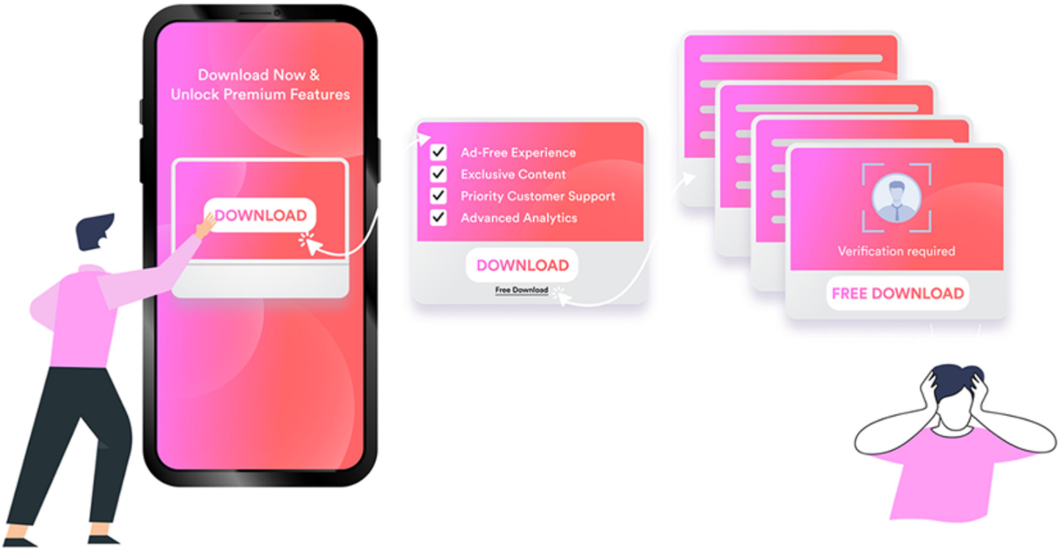
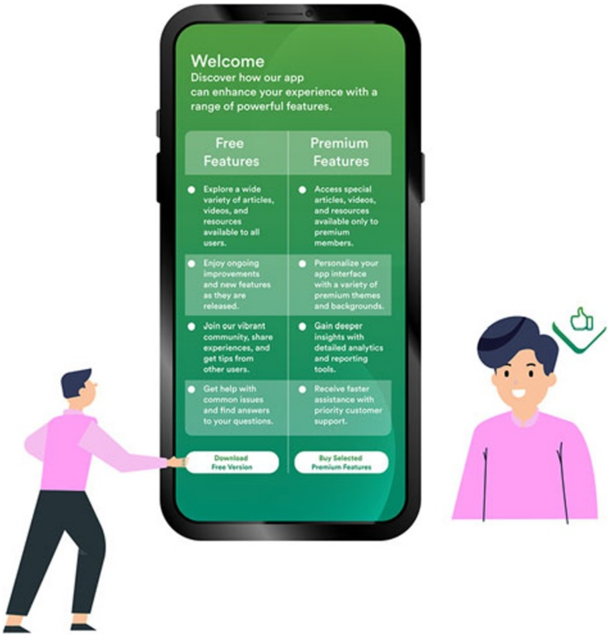


Figure 11. Unethical prompt designed bye fair patterns.

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

For example, an unethical prompt might instruct, ‘Create a call to action that encourages users to download the app and subscribe to the premium service’. This often leads to dark patterns where, upon clicking ‘Download’, users encounter a pop-up with preselected premium features and a paid subscription. The option for a ‘Free Download’ is hidden within multiple clicks, causing frustration and a sense of manipulation (Figure 11).

Conversely, an ethical prompt (Figure 12) might request, ‘Generate a clear and informative message explaining the app’s free features and benefits, with a separate option for users to access premium features and subscription details if they choose’. This results in a straightforward landing page explaining the app’s features, both free and premium, and a download button that clearly indicates whether the action is free or paid. Users can easily access the free download option without wading through confusing menus, and a transparent ‘opt-out’ for premium features is readily available.



Welcome
Discover how our app can enhance your experience with a range of powerful features.

Free Features	Premium Features
● Explore a wide variety of articles, videos, and resources available to all users.	● Access special articles, videos, and resources available only to premium members.
● Enjoy ongoing improvements and new features as they are released.	● Personalize your app interface with a variety of premium themes and backgrounds.
● Join our vibrant community, share experiences, and get tips from other users.	● Gain deeper insights with detailed analytics and reporting tools.
● Get help with common issues and find answers to your questions.	● Receive faster assistance with priority customer support.

Figure 12. Ethical prompt designed by fair patterns.

***

4. Conclusion

Through this article, we presented the way in which our digital interactions are heavily influenced by the use of dark patterns. These deceptive and manipulative design techniques are highly effective, offering short-term economic incentives and immediate advantages to their creators. However, their use harms consumers, who are forced to sacrifice their privacy by accepting obscure privacy policies or entering into contracts they had no intention of agreeing to. The subtle manipulation of decision-making environments in the digital space cannot be justified, as platforms often claim, by the promise of a high-quality customer experience. We no longer purchase – we are extorted. We no longer consent – we are violated. We no longer share – we are spied upon. The philosophy of dark patterns is ‘do as I tell you, not as you think’. Merchants, and sometimes even non-merchants, have industrialized manipulation processes, just as they once industrialized production processes. While European laws have made notable strides toward regulating dark patterns, a comprehensive solution to the problem remains elusive. The legislative focus has traditionally been on legal and technical

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

responses, overlooking the critical role of design in shaping user experiences. Yet, it is precisely within the design of interfaces – the space where users interact with technology – that significant influence is exerted, guiding or pressuring users in subtle ways. This gap highlights the need for a more integrated approach where design becomes a regulatory tool alongside traditional legal measures.

The concept of 'fairness by design', as proposed by the UK's Competition and Markets Authority, points to a promising direction. By acknowledging human cognitive limitations and the powerful impact of design on decision-making, this approach promotes design as a solution to mitigate coercion and manipulation. Initiatives like the Fair Patterns developed by Amurabi's Research & Development Lab further demonstrate how design can empower users rather than simply nudge them. These prototypes aim to foster autonomy by enabling users to make informed choices through accessible, clear, and fair information. Such design-centric strategies could be instrumental in the future of dark patterns regulation, where user empowerment becomes a regulatory objective. Moving forward, a regulatory framework that integrates design ethics and the principles of Fairness by Design will be essential – not only to ensure compliance but also to foster a culture that prioritizes user autonomy and well-being. Rather than being a mere tool for functionality or commercial gain, digital design should serve as a means to empower users, reinforcing trust and transparency in every interaction. This shift calls for a collaborative effort among designers, policymakers, and researchers to establish new industry norms that place human dignity at the forefront.

The challenge ahead is not just mitigating harmful practices but fundamentally rethinking digital experiences to actively empower users, safeguard their rights, and enhance their decision-making autonomy. This raises important questions: How can we balance innovation with ethical responsibility in design? What role should designers, regulators, and users themselves play in shaping a more transparent and fair digital landscape? As technology continues to evolve, an open and interdisciplinary dialogue will be crucial to ensure that digital environments serve not just business interests, but also the broader principles of trust, fairness, and human dignity.

Funding statement. None declared.

Competing interests. Marie Potel-Saville is CEO at FairPatterns & Amurabi.

References

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. *Proceedings of the 5th ACM Conference on Electronic Commerce*, 21–29.

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L., Komanduri, S., … Wilson, S. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, *50*(3), 1–41.

Acquisti, A., & Grossklags, J. (2007). What can behavioral economics teach us about privacy?. In A. Acquisti, (Ed.). *Digital Privacy: Theory, Technologies and Practices* (pp. 363–377). Boca Raton: CRC Press.

Agnew, W., Bergman, S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., and McKee, K. R. (2024). The Illusion of Artificial Inclusion. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–12.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Bacache-Beauvallet, M., & Bourreau, M. (2022). *Économie des plateformes*. Paris: La Découverte.

Bar-Gill, O. (2012). *Seduction by Contract: Law, Economics and Psychology in Consumer Markets*. Oxford: Oxford University Press.

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., & Lenzini, G. (2021). « I am definitely manipulated, even when i am aware of it. it's ridiculous ! » – dark patterns from the end-user perspective. *Designing Interactive Systems Conference 2021*, 763–776.

Citton, Y. (Ed.). (2014). *L'économie de l'attention: Nouvel horizon du capitalisme*. Paris: ?. La Découverte.

CNIL (LINC). (2019). *La forme des choix*. Cahiers IP Innovation & Prospective, 6. Accessed July 16, 2025. https://www.cnil.fr/sites/cnil/files/atoms/files/cnil_cahiers_ip6.pdf

Coase, R. H. (1937). The nature of the firm. *Economica*, *4*(16), 386–405.

Competition and Markets Authority (UK). (2022). *Online choice architecture: How digital design can harm competition and consumers*. Accessed July 16, 2025. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1066524/Online_choice_architecture_discussion_paper.pdf0

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

Dahlman, C. J. (1979). The problem of externality. *The Journal of Law & Economics*, *22*(1), 141–162.

Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020). UI dark patterns and where to find them: A study on mobile applications and user perception. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

European Commission, Directorate-General for Justice and Consumers. (2022). Behavioural study on unfair commercial practices in the digital environment – Dark patterns and manipulative personalisation – Final report. Publications Office of the European Union. https://data.europa.eu/doi/10.2838/859030

European Commission. (2024). Digital fairness – Fitness check on EU consumer law. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13413-Digital-fairness-fitness-check-on-EU-consumer-law_en

Faure, M. G., & Luth, H. A. (2011). Behavioural economics in unfair contract terms cautions and considerations. *Journal of Consumer Policy*, *34*(3), 337–358.

Graßl, P., Schraffenberger, H., Borgesius, F., & Buijzen, M. (2021). Dark and bright patterns in cookie consent requests. *Journal of Digital Social Research*, *3*(1), 1–38.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Gray, C., Santos, C., & Bielova, N. (2023). Towards a preliminary ontology of dark patterns knowledge. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–9.

Gunawan, J., Pradeep, A., Choffnes, D., Hartzog, W., & Wilson, C. (2021). A comparative study of dark patterns across mobile and web modalities. *Proceedings of the ACM 2021 Conference on Computer-Supported Cooperative Work and Social Computing*, 1–29.

Hosseini, H. (2003). The arrival of behavioral economics: From michigan, or the carnegie school in the 1950s and the early 1960s. *Journal of Socio-Economics*, *32*(4), 391–409.

International Organization for Standardization. (2019). *Ergonomics of human-system interaction - Part 210: Human-centered design for interactive systems*. Accessed July 16, 2025. (ISO 9241-210:2019). https://www.iso.org/fr/standard/77520.html

Jarovsky, L. (2022). Dark patterns in personal data collection: Definition, taxonomy and lawfulness. *Ssrn*, 1–51.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kaptein, M. (2015). *Persuasion profiling: How the internet knows what makes you tick* Amsterdam: Business Contact.

Khojah, R., Mohamad, M., Leitner, P., & Gomes de Oliveira Neto, F. (2024) Beyond code generation: An observational study of ChatGPT usage in software engineering practice. *Proceedings of ACM 2024 Software Engineering*, FSE 81(1), 1819–1840.

Kosch, T., & Feger, S. (2024). Risk or chance? large language models and reproducibility in Human-Computer Interaction research. *Interactions*, *31*(6), 44–49.

Krauss, V., McGill, M., Kosch, T., Thiel, Y. M., Schön, D., & Gugenheimer, J. (2024) "Create a fear of missing out" — ChatGPT implements unsolicited deceptive designs in generated websites without warning (draft), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Maier, M., & Harr, R. (2020). Dark design patterns: An end-user perspective. *Human Technology*, *16*(2), 170–199.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Marciano, A., Nicita, A., & Ramello, G. B. (2020). Big data and big techs: Understanding the value of information in platform capitalism. *European Journal of Law and Economics*, *50*(3), 345–358.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM 2019 Conference on Computer-Supported Cooperative Work and Social Computing*, 81(3), 1–32.

Mathur, A., Mayer, J., & Kshirsagar, M. (2021). What makes a Dark Pattern.. Dark? Design attributes, normative considerations, and measurement methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems: Making Waves, Combining Strengths*, 1–27.

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *A Journal of Law and Policy for the Information Society*, *4*(3), 543–565.

Miller, K., Sahni, N. S., & Strulov-Shlain, A. (2022). Sophisticated consumers with inertia: Long-term implications from a large-scale field experiment. *Becker Friedman Institute - Working Paper*, 1–67.

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark Patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Organisation for Economic Cooperation and Development (OECD). (2022). *Dark Commercial Patterns* , OECD Digital Economy Working Papers. 336. Accessed July16, 2025. https://www.oecd-ilibrary.org/science-and-technology/dark-commercial-patterns_44f5e846-en

Pareto, V. (2014). *Manual of Political Economy: A Critical and Variorum Edition*. A. Montesano & al. Oxford: Oxford University Press.

Potel-Saville, M., & Da Rocha Francois, M. (2023). From dark patterns to fair patterns? usable taxonomy to contribute solving the issue with countermeasures. *Proceedings of the 2023 Annual Privacy Forum*.

https://doi.org/10.1017/cfl.2025.10018 Published online by Cambridge University Press

Rasch, A., Thöne, M., & Wenzel, T. (2020). Drip pricing and its regulation: Experimental evidence. *Journal of Economic Behavior & Organization*, *176*, 353–370.

Rochelandet, F. (2010). *Économie des données personnelles et de la vie privée*. Paris: La Découverte.

Schäfer, H.-B., & Leyens, P. (2010). Judicial control of standard terms and european private law. In P. Larouche, and F. Chirico (Eds.), *Economic Analysis of the DCFR: The Work of the Economic Impact Group within the CoPECL Network of Excellence* (pp. 99–121). Munich: Sellier European Law Publishers.

Simon, H. A. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Macmillan.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *American Economic Review*, *49*(3), 253–283.

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, Communications and the Public Interest* (pp. 37–72). Baltimore: John Hopkins Press.

Smythe, D. W. (1950). Television and its educational implications. *Elementary English*, *27*(Jan), 41–52.

Stigler Center. (2019). *Report of the Committee for the Study of Digital Platforms*. Accessed July 16, 2025. Chicago: University of Chicago. https://www.chicagobooth.edu/-/media/research/stigler/pdfs/digital-platforms—committee-report—stigler-center.pdf

Sunstein, C. (2014). Choosing not to choose. *Duke Law Journal*, *64*(1), 1–52.

Sunstein, C., & Thaler, R. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, *70*(4), 1159–1202.

Susser, D., Roessler, B., & Nissenbaum, H. F. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, *4*(1), 1–45.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth and happiness*. New Haven: Yale University Press.

Thaler, R., & Sunstein, C. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.

Vaswani, A. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Verhelst, E. W. (2012). *Recht Doen aan Privacyverklaringen: Een Juridische Analyse van Privacyverklaringen op Internet*. Deventer: Wolters Kluwer.

Vial, S. (2021). *Le Design*. Paris: Presses Universitaires de France.

Voigt, C., Schlögl, S., & Groth, A. (2021). Dark patterns in online shopping: of sneaky tricks, perceived annoyance and respective brand trust. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 143–155.

Wagner, G. (2010). Mandatory contract law: Functions and principles in light of the proposal for a directive on consumer rights. *Erasmus Law Review*, *3*(1), 47–71.

Waldman, A. E. (2020). Cognitive biases, dark patterns, and the "privacy paradox. *Current Opinion in Psychology*, *31*, 105–109.

Warin, T., & Leiter, D. (2012). Homogeneous goods markets: An empirical study of price dispersion on the internet. *International Journal of Economics and Business Research*, *4*(5), 514–529.

Warin, T., & Troadec, A. (2016). Price strategies in a big data world. *Encyclopedia of E-Commerce Development, Implementation, and Management 1*, 625–638.

Yeung, K. (2017). "Hypernudge": Big data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136.

Zachrisson, J., Storrø,G., & Boks, C. (2012). Using a guide to select design strategies for behaviour change: theory vs. practice. In *Proceedings of EcoDesign 2011 (30 November–2 December, Kyoto, Japan)*, 362–367.

Zuboff, S. (2020). *L'Âge du capitalisme de surveillance. Intérêts et enjeux*. Paris: Zulma.

Zuiderveen Borgesius, F. (2015). *Improving Privacy Protection in the Area of Behavioural Targeting*. Deventer: Wolters Kluwer.

Fabien Lechevalier is a PhD candidate in law at *Paris-Saclay University*. He is a researcher at the *Centre d'Études et de Recherche en Droit de l'Immatériel (CERDI)* and an affiliate of Stanford University's *Transatlantic Technology Law Forum*. His research focuses on the collective dimension of privacy, data governance models, and legal design. He was a visiting fellow at Cornell Tech's *Digital Life Initiative*, and currently teaches law at several universities (*Paris-Saclay University, Mines-Télécom Institute*, etc.).

Marie Potel-Saville is the founder of *Amurabi*, a legal innovation studio using design, and of www.fairpatterns.com, a platform dedicated to combating dark patterns. She serves on the Advisory Board of the *Legal Lab* at the Serpentine Gallery and the consulting firm *Pickering Pierce*. Today, she is a keynote speaker at numerous international conferences (*IAPP Global Privacy Summit, W@Privacy*, etc.) and teaches legal innovation through design at several universities (*SciencesPo Paris, Singapore Management University*, etc.).

Cite this article: Lechevalier F. and Potel Saville M. (2025). Fairness by design: Combatting deceptive AI-driven interfaces. *Cambridge Forum on AI: Law and Governance* **1**, e31, 1–20. https://doi.org/10.1017/cfl.2025.10018