Biology & Philosophy (2022) 37:44
https://doi.org/10.1007/s10539-022-09872-z

The argument from Evel (Knievel): daredevils and the free energy principle

Sidney Carls-Diamante[1]

Received: 25 November 2020 / Accepted: 14 August 2022 / Published online: 11 September 2022

Abstract
Much of the literature on the *free energy principle (FEP)* has focused on how organisms maintain homeostasis amidst a constantly changing environment. A fundamental feature of the FEP is that biological entities are "hard-wired" towards self-preservation.

However, contrary to this notion, there do exist organisms that appear to seek out rather than avoid conditions that pose an elevated risk of serious injury or death, thereby jeopardizing their physiological integrity. Borrowing a term used in 1990s popular culture to refer to stunt performers like Evel Knievel, these organisms that exhibit such behavioural characteristics can be referred to as *daredevils.*

This paper presents the case of daredevils as a challenge to the FEP's homeostasis- and optimization-based construal of biological systems. It also introduces three possible explanatory strategies by which the FEP can account for daredevils. The broader objective of the paper is to enhance the FEP's ability to account for a diverse range of complex behaviour.

Keywords Free energy principle (FEP) · Self-preservation · Dark room problem · Daredevils · Thrill-seeking

Introduction

The *free energy principle (FEP)* is a wide-ranging theory of life whose scope includes brain function and evolution (Colombo and Wright 2018; Friston 2009; Hohwy 2015; Ramstead et al. 2019). It has been described as a "theoretical framework for addressing the self-organization in biological systems, focusing on the interdependency of

✉ Sidney Carls-Diamante
sidney.carls-diamante@uni-konstanz.de

[1] Zukunftskolleg/Philosophy Department, University of Konstanz, Universitätsstrasse 10, 78464 Konstanz, Germany

brain, body, and local environment" (Kirchhoff 2018, 2524). One of its central claims is that the deep structure of brain function reflects a causal structure fundamental to biological systems from the molecular to the organism level of organization (Friston 2009, 2010; Hesp et al. 2019; Ramstead et al. 2019). This causal structure, referred to as *free energy minimization* (to be defined and discussed in more detail in the next section), can be operationalized into the notion that biological entities self-organize their components in order to maintain homeostasis or physiological stability. Minimizing free energy can also be described as a biological entity's means of maintaining its internal order (see Schreiber and Gimbel 2010). Importantly, this notion is presented as a universal claim: free energy minimization has been posited as a "necessary, if not sufficient biological characteristic" (Friston et al. 2006, 74) and even a "biological imperative" (Friston et al. 2006; Friston and Stephan 2007). FEP theorists also hold that "the defining characteristic of biological systems is that they maintain their states and form in the face of a constantly changing environment," such that "biological agents, like animals or [brains] resist a natural tendency to disorder" (Friston 2010, 127). This notion that biological entities minimize free energy in order to maintain their physiological integrity and avoid decay is a claim about the causal organization biological entities evolved to exhibit, which in turn distinguishes biological from non-biological entities (Friston 2010; Varela et al. 1974).

These features of the FEP thus entail that living organisms are characterized by the tendency towards self-preservation (Kirchhoff 2018), in order to avoid decay or physiological destabilization (Colombo and Wright 2018; Hesp et al. 2019). Indeed, the biological imperative has been articulated as "avoid surprises and you will last longer" (Friston et al. 2012, 2). One self-preservation strategy is *selective exposure*, or limiting exposure to environmental conditions that are familiar and beneficial or at least not harmful to the organism. By keeping itself within familiar conditions, the organism can anticipate future outcomes with greater accuracy—i.e., it is less likely to encounter surprise—and thus is more likely to respond to them more optimally than would have been the case in unfamiliar conditions. Conversely, unfamiliar conditions are more likely to be surprising, and thus pose a higher risk of being deleterious to the organism.

Although it is one of the FEP's fundamental claims, the biological imperative to "avoid surprises" renders the theory vulnerable to counterexamples. Particularly, organisms do exist that actively and deliberately expose themselves to deleterious environmental conditions, thereby risking injury or death. For a subset of these organisms, this is a result of species-specific stereotypic behaviour. For instance, some male praying mantises and spiders go on to mate with females despite the risk of being cannibalized as part of the mating process (Lelito and Brown 2006; Prokop and Václav 2005). In some species of exploding ants, "during territorial combat, workers…sacrifice themselves by rupturing their gaster and releasing sticky and irritant contents of their…mandibular gland reservoirs to kill or repel rivals" (Laciny et al. 2018, 1–2). In other cases, organisms injure themselves in order to avoid potentially greater future danger, such as trapped deer chewing off an ensnared leg (Ramsden and Wilson 2014). Perhaps the most famous example of self-injurious behaviour is that of lemmings, whose rushes off cliffs into the sea have long been misinterpreted as suicide, but which scientific investigations have subsequently revealed to

be mass emigrations out of an overpopulated and resource-depleted habitat into a less crowded one (Chitty 1996).

However, this paper is more interested in cases wherein self-exposure to potentially deleterious conditions initially appears to have no immediate biological benefits and is not typical of the species. In particular, I have in mind humans who engage in acts that are well-known to be risky and dangerous, such as eating laundry detergent pods, holding a glass around one's mouth and sucking the air out to create a vacuum with the goal of making one's lips swell, dousing oneself in flammable liquid and setting oneself alight, or rapidly ingesting a spoonful of ground cinnamon. If these behaviours sound familiar, it is because they are all examples of "challenges" gone viral on social media. Individuals who perform these acts will be referred to as *daredevils.*

The defining characteristic of daredevils is that they actively seek out environmental conditions that present a heightened risk of injury, impairment or death. Actions that fit this description—such as those enumerated in the previous paragraph—will be called *daredevil behaviour.* Furthermore, in contrast to the insects and spiders, deer, and lemmings mentioned above, daredevil behaviour does not appear to be the result of evolved, inherited, or species-specific behavioural patterns generalizable to humans. Nevertheless, it has been observed that teenagers or young adults are more likely to engage in daredevil and other forms of risky behaviour than their older counterparts (Brodbeck et al. 2012; Caffray and Schneider 2000); however, the influence of developmental stages on predispositions towards risky behaviour is a matter that requires independent investigation, beyond what this paper can provide.[1] Daredevils pose a challenge to the FEP because they are putative real-world counterexamples to its biological imperative and thus construal of the behaviour of living organisms, as they deliberately approach rather than avoid potentially deleterious environmental conditions. In other words, rather than avoid surprises, they appear to intentionally seek them out.

The FEP extensively accounts for how organisms maintain their homeostasis amidst a changing environment (Hesp et al. 2019), thus fulfilling a hard requirement for any theory of life. It has been said that the FEP's viability hinges on its universality (Hohwy 2015; Kiverstein 2018). Consequently, in order for it to achieve explanatory completeness, it must also account for organisms and behaviour that appear to contradict its claim about the fundamental nature of living organisms. In light of these factors, this paper explores the nature of daredevils and the challenge it poses to the FEP via the biological imperative.

The rest of the paper proceeds as follows. Section "The free energy principle" provides an exposition of the FEP in Sect. "The free energy principle". The dark room problem is presented in Sect. "The dark room problem: from the safe side…", and daredevils and the challenge they pose are introduced in Sect. "…to the wild side: daredevils". Three potential strategies for resolving the problem of daredevils are proffered in Sect. "Challenge accepted". At this point, it must be noted that these strategies are tentative solutions rather than completed accounts. They are intended

[1] Age and developmental stages are significant factors that need to be taken into account by theories of human behaviour. Thus far, however, patterns in adolescent and early adult behaviour (of which risk-taking is notable) appear to be an area that has received little attention from the FEP.

as conversation starters to draw attention to risky or maladaptive behaviour, which thus far have remained a gap in FEP literature, in the hope that more research will be dedicated to these topics in the future. A brief conclusion in Sect. "Concluding remarks" closes the paper.

The free energy principle

Free energy is a term used in thermodynamics and statistical physics, albeit with varying senses (Hesp et al. 2019). While the FEP is concerned with free energy in the context of statistical physics (Friston 2010), understanding free energy in a thermodynamic context is nevertheless helpful.

Thermodynamic free energy is a measure of useful energy (Schreiber and Gimbel 2010), more specifically the "measure of energy available to do useful work" (Kirchhoff 2018, 2523). When energy is converted from its potential to kinetic form, it becomes more disordered in the process; thus, *entropy*, or the amount of disorder, increases. The more disordered energy becomes, the more difficult it becomes to "extract it to do work" (Kirchhoff 2018, 2523): as thermodynamic free energy—i.e., ordered energy that can be used to do work—decreases, entropy increases. As such, the efficiency of a system depends on how well it can avoid disorder or entropy in the process of converting potential energy to kinetic energy.

A conceptual switch is necessary in order to understand free energy in the FEP context. The free energy in question at this point is *variational*, rather than thermodynamic. The FEP's proponents define variational free energy—or simply free energy—as a statistical probability distribution that "bounds surprise, conceived as the difference between an organism's predictions about its sensory inputs…and the sensations it actually encounters (Friston et al. 2012, 1). Free energy is thus a "measure of the difference between what the organism senses and what it expects to sense" (Hesp et al. 2019, 196). Within the FEP, free energy is conceptually parallel rather than contrary to entropy, such that "minimizing free energy amounts to minimizing entropy" (Kirchhoff 2018, 2523). While this reading identifies free energy as a statistical quantity rather than a physical state, it can be operationalized as a sensory state experienced by an organism that differs from the sensory conditions it has anticipated through the use of an internal model of the environment (which will be discussed in more detail shortly).

Since an organism is more likely to respond sub-optimally to an unanticipated or *surprising* sensory state than it is to one that corresponds to the conditions it has anticipated, surprising states are potentially deleterious. Now, it must be noted that due to phenotypic diversity, surprising states differ across individuals, even within the same species. A sub-optimal response to a surprising state increases the likelihood of injury, impairment, or death, which result in drastic physiological alterations and must therefore, according to the FEP, be avoided. These changes are often referred to as *phase transitions*, as they involve significant modifications to the configuration or phase of the organism's components. Thus, the FEP's imperative to "avoid surprises and you will last longer" (Friston et al. 2012, 2) is a directive for self-preservation by keeping away from potentially deleterious conditions as much as possible. The rela-

tionship between free energy and surprise can be summarized as such: "free energy is a *proxy* for a quantity called surprise…, which reflects the improbability of finding an organism in some sensory state (Ramstead et al. 2019, 190). Free energy can be proximal as well as cumulative. The former pertains to free energy related to surprising states in the immediate or proximal future. The latter refers to free energy accumulated over the course of the organism's lifespan: the more an organism finds itself in surprising states, the more free energy it will accumulate, thereby heightening its risk of undergoing phase transitions.

Another important aspect of the FEP is the notion that biological organisms possess a *generative model*, or internal model of the world that is "implicit in [their] phenotype" (Friston et al. 2006, 78), and which "[encodes] the statistical structure of their local environment" (Ramstead et al. 2019, 190). The generative model is constructed from bottom-up signals received via the organism's sensorium, which must then be "combined" in order to recreate what the world is like as correctly as possible. That is, the generative model *infers* or forms hypotheses about how to reconstruct the causal matrix that is responsible for the structure of the world as it is (Clark 2013). In doing so, the generative model draws on stored information, known as *prior beliefs* (or simply *priors*), to support its inferences. Priors can either be innate, or encoded in the organism's phenotype (Ramstead et al. 2019), or learned through experience over the course of the organism's lifespan. As such, the generative model is the organism's main source of information about the world.

Another crucial function of the generative model is that it generates predictions or expectations about future states the organism might experience (Hesp et al. 2019), based on priors and information presently received via the sensorium. Like ripples in a pond that radiate away and back to the object that disturbed the water, an organism's interactions with the environment result in changes to its external surroundings and internal sensory states. To a certain extent, the organism is able to predict what these changes will be like. However, since the environment is in a constant state of flux, the actual external conditions that arise and the corresponding sensory states the organism experiences as a consequence of these external states do not always match the organism's expectations (Kiverstein 2018). Since the accuracy of the predictions is inversely proportional to the amount of free energy generated, it is in the best interests of the organism to bring the model up to a satisfactory level of accuracy that allows it to effectively minimize free energy.

Optimization of the model and minimization of free energy involves changing "two quantities on which free energy depends" (Kirchhoff 2018, 2526): the external environment and the organism's internal states. The environment is modified through action, and the internal model through perception (Friston 2010). An organism can act on its surroundings in order to bring their configuration as close as it can to a state that generates sensory signals that conform as accurately as possible to its expectations. Meanwhile, perceptual experiences produce sensory signals that are incorporated into the internal model, thereby updating it. These signals reinforce or revise the information stored in the model, depending on whether they conform or diverge from it.

The FEP takes the generative model a step further. It claims that a biological organism is an embodied and *implicit* model of its own existence, such that it "is

defined by the [physiological] states that it needs to maintain within certain bounds if it is to continue to exist" (Kiverstein 2018, 563). Because phenotype determines an organism's physiology, behavioural patterns, cognitive capacities, and environmental context, it is likewise instrumental to delimiting the states the organism can and will experience throughout the course of its life. Phenotype also specifies what constitutes homeostasis for an organism, as well as the regulatory processes necessary for that organism to maintain homeostasis (Kiverstein 2018). Importantly, however, phenotypic diversity within any given species entails a degree of variation between what is homeostatic or surprising for the individual. While many homeostasis-maintaining processes are metabolic, others are behavioural: organisms tend to seek out conditions that are amenable to keeping themselves within homeostasis, a tendency known as selective exposure. The FEP provides the following example, of "an insect that 'prefers' the dark; imagine an insect that evolved to expect the world is dark. It will therefore move into shadows to ensure it always samples a dark environment" (Friston et al. 2006, 78), where sampling refers to receiving sensory input consistent with the dark environment, thus meeting the insect's expectations.

Since biological entities—especially organisms—are situated within ecological niches, (Friston et al. 2006; Ramstead et al. 2019), they are more likely to experience certain environmental conditions than others. This is because the number of elements within any given niche is finite relative to the broader environment, and as such can enter into a limited number of configurations. When the organism limits its exposure to a particular niche—especially a familiar one—the set of possible conditions that it can expect to experience is significantly narrowed down. For instance, to an octopus in the wild, the probability of coming face to face with a shark is much higher than the probability of being sideswiped by a strange man on a motorcycle. Probability distributions corresponding to the likelihood of experiencing certain ecological conditions more than others are operationalized as the states that the organism "*expects* given its phenotype and the eco-niche it lives in" (Kiverstein 2018, 563, italics in original). These probability distributions are set in place and reinforced through different mechanisms, like exposure to the ecological niche in question, or information encoded in the phenotype and hereditarily (and possibly epigenetically) transmitted (Friston et al. 2006, 2012; Hesp et al. 2019). As such, the generative model encodes the repertoire of physiological states and ecological conditions that the organism can normally expect to experience throughout its lifetime, based on its phenotypic profile *and* individual circumstances (Ramstead et al. 2019). This information then becomes the source of the system's "implicit beliefs about the outer world" (Hesp et al. 2019, 196), which in turn are recruited for behaviour control.

The dark room problem: from the safe side…

The FEP's biological imperative has previously been challenged, in a prominent thought experiment known as the *dark room problem* (Friston et al. 2012). The dark room problem runs as follows: if it is true that biological entities are characterized by the imperative for self-preservation via free energy minimization, operationalized as avoiding surprising states, why do living organisms not stay in dark rooms all their

lives? In contrast to any given "normal" environment, which is in a constant state of flux, nothing can generate conditions as uniform and unsurprising as a room that is dark at all times. Nevertheless, *actual* living organisms do not confine themselves to dark rooms, but move around in the world, surprises and all. The fact that living organisms "do not seem to avoid surprises" (Friston et al. 2012, 1) appears to directly contradict the FEP's claim that biological entities are hard wired to minimize free energy and thus avoid unanticipated and hence potentially deleterious states.

FEP theorists respond that unless the organism in question is one whose natural environmental niche is a consistently dark room, staying in one all its life is actually maladaptive and deleterious (Friston et al. 2006, 2012). The canonical FEP explanation is that "a dark room will afford low levels of surprise if, and only if, the agent has been optimized by evolution (or neurodevelopment) to predict and inhabit it" (Friston et al. 2012, 3). On the other hand, an organism that is not endemic to a dark room will embody a model of an environment that is rich in sensory stimuli. For this organism, a dark room will not be part of its repertoire of expected states, and will thus prove surprising were the organism to find itself in such conditions. In order to minimize free energy, the organism "will leave at the earliest opportunity" (Friston et al. 2012, 3), to seek out a more stimulating environment that is in line with its expectations. Thus, FEP theorists' resolution of the dark room problem demonstrates that immersion in an environment devoid of sensory stimuli does not necessarily equate to minimizing free energy. Rather, the states that are surprising and unsurprising to an organism depend on the kind of model of the world it has.

I have discussed the dark room problem here largely for historical reasons, but also because like daredevils, the dark room problem illustrates how real-world observable behaviour of living organisms differs from the FEP's construal of the nature of biological systems. Moreover, the dark room problem and daredevils demonstrate the extreme opposite ends of the spectrum of free energy-minimizing behaviour. As such, they can be thought of as counterweights to one another: the counterfactual organisms posited in the dark room problem exhibit what is putatively extreme surprise avoidance, while daredevils are characterized by what appears to be extreme surprise seeking. Thus, the dark room problem attempts to refute the FEP's exhortation to avoid surprises, whereas daredevils challenge its self-preservation imperative. The discussion now proceeds to examining what can be called the *daredevil challenge* in more detail.

…to the wild side: daredevils

The daredevil challenge is shorthand for the following problem: If biological entities are hard-wired for self-preservation, in so doing avoiding conditions that jeopardize their homeostasis, then why do some organisms—i.e., daredevils—deliberately place themselves in potentially deleterious situations? Taking inspiration from professional stunt performers such as Evel Knievel—who is famous for his motorcycle jumps over several rows of cars, buses, wild animals, or even a canyon (The Editors of Encyclopaedia Britannica, 2020)—daredevils are individuals (in this case certain

humans) who perform extremely risky actions, i.e., *daredevil behaviour*, contrary to those associated with self-preservation.

The daredevils of present interest are known for their thrill-seeking, risk-taking behaviour—sometimes designated in the literature as *T behaviour* (Morehouse et al. 1990; Sarshar et al. 2019; Self et al. 2007). Modern psychology sometimes refers to individuals prone to such behaviour as having a *Type T personality* (Morehouse et al. 1990), whose common manifestations are engaging in extreme sports (Self et al. 2007), anti-social behaviour, unsafe sexual practices, or recklessness in everyday activities. Important features of daredevil behaviour are that (1) it is not prompted by biological requirements (e.g., a desperate need for food) or threatening environmental conditions (e.g., to escape predators, responses to territorial conflict); (2) are not part of the phenotypically encoded "normal" behavioural repertoire (e.g., mating behaviour) of the species as a whole but exhibited only by certain individuals; (3) pose an extremely elevated risk of injury or death; (4) and do not appear to have biological benefits. Instead, it appears that "motivation for much risk-taking and engagement with uncertainty is simply the thrill of it" (Sarshar et al. 2019, 1). With social media in full flower, it is not difficult to come across daredevils: one need only log on to social media and do a cursory search to find videos of individuals engaged in dangerous activities such as ingesting laundry detergent capsules, setting oneself on fire as a prank, or practicing yoga or acrobatics on the balcony railing of a high-rise apartment or even a cliff edge.

Significantly, unlike Knievel and other professional stunt performers who are highly trained, the daredevils I have in mind are not practiced hands at the daredevil behaviour in question (in the case of "extreme yoga," proficiency in yoga is quite different from practicing it on precarious surfaces). Evidence for the surprising nature of daredevil behaviour comes from the daredevils' own admissions to not having previously tried doing the actions in question (which they often admit in the videos concerned), their astonished, shocked, or downright terrified reactions to the immediate consequences of their actions, and at times the mere fact that they are still alive and uninjured and thus able to engage in the daredevil behaviour concerned in the first place. Thus, another important characteristic of daredevil behaviour is that 5) it is novel.

Daredevils come in three types. First, there are those individuals who follow trends or others' behaviour that they have seen, usually on social media, for the very first time. The second type are those who engage in daredevil behaviour without being prompted to. Finally, the third type refers to those who engage in progressively risky behaviour; in this case, danger is not necessarily due to the kind of behaviour per se, but the elimination of safety precautions that normally accompany it. The challenge to the FEP posed by the first two types of daredevils pertains to action selection, i.e., why agents would seek out highly surprising and potentially deleterious states, whereas the third type raises the question of why agents persist in this type of action selection.

The FEP claims to be a universal theory about the fundamental structure of biological entities, applicable at all levels (Friston 2009; Hohwy 2015; Kiverstein 2018; Limanowski and Blankenburg 2013). It has been written of the FEP it must be accepted in its entirety or not at all, the reason being that what the FEP makes are

overarching claims about the fundamental nature of all biological entities (Hohwy 2015). To make this point, Hohwy (2015) compares the FEP to evolutionary theory. He argues that for it to have genuine explanatory value, a theory of evolution must be applicable to all species since its purpose is to identify a unifying, general principle that holds true regardless of the extent of biological diversity. In the same vein, for the FEP to fulfil its goal of accounting for all biological phenomena using the same explanatory toolkit, its claims must be *universally* true. As such, the FEP must also be capable of accounting for behaviour that appears to be contradictory to its self-preservation mandate.

Furthermore, the fact that daredevils and Type T personalities are recognized as a category implies that they are a subclass of humans rather than isolated or anomalous occurrences (which the FEP would nevertheless also have to account for). Thus, daredevils and daredevil behaviour need to be firmly ensconced within the FEP's explanatory framework, instead of standing out as counterexamples or special cases. That is to say, the FEP must have a principled account of why daredevil behaviour occurs: without it, the FEP's explanatory force and scope would be jeopardized.

The five characteristics of daredevil behaviour enumerated above seem to be in disharmony with the FEP's directive to avoid potentially deleterious surprising states. That daredevil behaviour is not typical of humans as a species further implies that it may be due to "atypical" neural or psychological features, and thus warrant closer investigation. Since daredevil behaviour is not due to strict biological or ecological requirements, poses an increased likelihood of phase transitions, and is highly surprising, it can be safely described as suboptimal. As such, daredevils and daredevil behaviour embody the following question, which challenges the FEP's standard characterization of living organisms: "If the brain is built to make optimal decisions, then why does it make so many suboptimal ones?" (Rahnev and Denison 2018, 17). Daredevils thus pose a potential threat to theories that place considerable emphasis on optimality, such as the FEP.

It is in light of this need to account for daredevil behaviour—i.e., to address the daredevil challenge—that this paper proffers a number of solutions that can be developed further in future FEP research. These solutions will be discussed in the following section. There are two caveats that must be kept in mind regarding the succeeding discussions. First, the starting points for the solutions are extant claims and explanatory strategies in the FEP corpus, particularly regarding prior beliefs, epistemic behaviour, the generative model, and the role of dopamine. Consequently, the arguments below are reflective of the state and lines of reasoning of the FEP literature. Second, the broad objective of presenting these solutions—and this paper, for that matter—is, so to speak, to get the discussion ball rolling on risky and maladaptive behaviour in light of the FEP. Thus, they serve as conversation starters rather than completed accounts (which a single paper would be hard put to provide). Consequently, the solutions presented here can be developed further in future FEP research.[2]

[2] Or replaced in favour of more viable accounts when the time comes.

Challenge accepted

As mentioned above, the first two types of daredevils raise issues about initial action selection, while the third is concerned with persistent action selection. In the same vein, the first two solutions focus on why daredevil action policies are selected, while the third accounts for progressively risky behaviour. The three solutions have different aspects of daredevil behaviour and the FEP as their starting points, and thus vary in their explanatory strategies. They are, however, complementary to one another, and can be recruited or integrated by future research into a comprehensive account of daredevil behaviour.

Faulty priors

The first solution involves daredevils having "faulty" priors that provide unreliable information about the consequences or riskiness of the daredevil behaviour in question. To begin with, the FEP maintains that agents need to believe that their actions minimize free energy (Friston et al. 2015); this can be operationalized as the agent believing (with varying degrees of consciousness or awareness about the belief) that their actions will bring them biological, psychological, or social benefits. Following the FEP's line of reasoning, action selection is influenced by the belief that performing the action in question will be the most effective at minimizing free energy (Friston et al. 2015), i.e., generating benefits or avoiding harms and other negative states. Conversely, *action policies* or sequences of actions "that do not minimize expected free energy are a priori surprising and will be avoided" (Friston et al. 2015, 195). However, daredevils do what appears to be exactly the opposite: the action policies they select are highly surprising, and thus increase rather than minimize free energy and inflate the likelihood of undergoing the very phase transitions they are meant to avoid.

Now, most other humans know *a priori* that the outcomes of daredevil behaviour are likely to be deleterious, and thus avoid it. The FEP's claim that the agent selects the action policy that is the most effective at minimizing free energy implies that the daredevil believes and subsequently predicts that the outcomes of daredevil behaviour in question will be (1) beneficial and (2) may not be exceedingly risky (Caffray and Schneider 2000). Daredevil behaviour may be motivated by beliefs that it will enhance positive affective states (e.g., enjoyment, excitement, or thrill), reduce negative affective states (e.g., depression, stress, or boredom), or bring about social rewards (e.g., higher status in the community, acceptance into a peer group, asserting one's identity) (Brodbeck et al. 2012; Caffray and Schneider 2000). At this point, it must be noted that the perceived benefit differs according to the type of behaviour in question. Importantly, such beliefs about the benefits of daredevil behaviour can influence other judgments about the nature and dangerousness of the action: it has been found that "specific cognitions may influence initiation into risky behaviours and that further involvement may thus result in subsequent changes in cognition" (Caffray and Schneider 2000, 547). A common consequence of what these authors refer to as "cognitive distortion" is that a person—especially a risk-prone one—may

underestimate the riskiness of daredevil behaviour, and thus become "overly optimistic about the consequences of their behaviour" (Caffray and Schneider 2000, 547).

When these factors are taken together with the fact that what is surprising differs significantly across individuals, a reasonable hypothesis is that daredevils have prior beliefs about the nature and consequences of daredevil behaviour that differ from those of their non-daredevil peers. In particular, a daredevil's risk assessment may be unreliable and hence faulty (Caffray and Schneider 2000; Rahnev and Denison 2018), such that they are overly optimistic or erroneously evaluate the hazards and potential danger of daredevil behaviour. As such, their predictions about the outcomes of daredevil behaviour are likely to be inaccurate when compared to those of their non-daredevil peers who may have a more accurate appraisal of the situation. In the same vein, daredevils' generative models may be flawed in reconstructing the causal matrix of the world, thus resulting in incorrect beliefs about the cause-and-effect relationships related to daredevil behaviour.

Another aspect to consider is the modality of experience of daredevil behaviour. Signals of different modalities provide dissimilar estimates of the phenomenon in question (Rahnev and Denison 2018). For this reason, the more multimodal the information accessed by the generative model, the more accurately it can formulate beliefs and predictions. In contrast, much of daredevil behaviour is copied off what one sees, often on social media; this is particularly true in the case of the first type of daredevil. Moreover, there is evidence that perceiving one's peers engaging in risky behaviour is a stronger predictor that one will follow suit than having previous first-hand experience of risky behaviour is (Caffray and Schneider 2000). Consequently, beliefs about the nature, experience, and consequences of the daredevil behaviour in question are formed predominantly on the basis of exteroceptive input. This information is largely visual, but also auditory. However, because the daredevil behaviour is at this point completely novel, stored and real-time interoceptive signals about it—which would have provided information about internal states related to danger perception such as increased heart rate, heightened blood pressure, or pain—may be conspicuously sparse. Significantly, interoceptive information is vital to maintaining homeostasis by "informing other neural systems about the internal state of the body" (Gu and FitzGerald 2014, 1), which in turn has extensive influence on decision-making. The sparseness of interoceptive signals that could have provided information about negative physiological states accompanying daredevil behaviour may likewise skew beliefs and consequent predictions to the effect that daredevil behaviour appears more pleasurable and less noxious or risky than it actually is, thereby contributing to inaccurate risk assessment. Furthermore, the "avoidance of atypical events" such as daredevil behaviour "seems more intuitive for physiological states (as reflected in interoception) than for environmental states (as signalled by exteroception)" (Seth 2014, 271). Thus, without robust interoceptive information, the daredevil's generative model registers daredevil behaviour as far less deleterious than it truly is, and is likelier to select an action policy in its favour (Friston et al. 2014, 2).

Another factor to consider is that emotional states are known to affect perceptual judgements (Rahnev and Denison 2018). It is known that emotional displays "[elicit] behavioural responses from others, the detection of which could serve to confirm predictions of interoceptive condition" (Seth 2013, 568), thereby reinforcing decisions

to engage in daredevil behaviour. A commonality between the videos of daredevil trends on social media is that the participants are usually in a high state of arousal (e.g., thrill, excitement, or giddiness) while performing the action. Moreover, in many videos, the individuals actually executing daredevil behaviour are surrounded by an "audience" egging them on, thereby intensifying emotional arousal. Taken together with the sparseness of interoceptive information that could have signalled negative sensations associated with risky actions, these emotional displays can reinforce the daredevil's existing beliefs and consequent predictions that the daredevil behaviour to be copied will be exciting or enjoyable, beneficial, and do not pose that high a threat of serious injury or even death.

Daredevil behaviour as epistemic behaviour

Still addressing the issue of action selection, let us take as an example a daredevil we will call Evel. Evel is riding his trusty motorcycle through unfamiliar territory, and sees a canyon for the very first time. After having admired the canyon's magnificent breadth and depth, he decides to jump across to the other side on his motorcycle, even though doing so poses an elevated risk of falling to his death—the ultimate phase transition the FEP cautions against.

This second explanation for daredevil behaviour understands it as a type of *epistemic behaviour*. Epistemic behaviour is a means of gaining new information about the world via action, in the process updating the generative model (Friston et al. 2017). Preface to the account at hand is the FEP's construal of agents as being "coupled to the environment through observation (sampled from the generative process) and actions (sampled from its posterior beliefs," and that "[to] couple the agent to its environment, we have to specify how its expectations depend upon observations and how its action depends upon expectations" (Friston et al. 2014, 3). Thus, when encountering a novel environment, the agent observes its features and subsequently formulates hypotheses about them. Some of these hypotheses would pertain to the roles the features observed play in the environment, as well as the possibilities for action that they afford. These hypotheses in turn form the bases of expectations or predictions about the states that the agent can encounter in the given environment. Since the environment is as yet novel, these hypotheses will not have been tested, and so pose considerable uncertainty. Likewise, expectations about the environment will not have been assigned statistical weightings corresponding to their plausibility.

Since the environment is novel and thus still highly surprising, and the FEP mandates that an agent must minimize uncertainty about its hypotheses by active sampling (Friston et al. 2017), Evel must select an action policy that will minimize uncertainty about the hypotheses the generative model is currently trying to test. The hypothesis, in this case, is whether *he, Evel*, can jump across the canyon on *his* motorcycle. The salient point in this hypothesis, i.e., the information that needs to be gained, is whether Evel—and not another person, or a theoretical possibility—can make the jump on his trusty motorcycle, which in virtue of frequent usage has become deeply encoded into his model of the world. Thus, information about whether a motorcycle jump of that length is possible for him and his motorcycle is an important way of refining Evel's model of the world. In addition to reducing the "first level of uncer-

tainty [which] is about the cause of sensory outcomes under a particular policy" (Friston et al. 2017, 2636), this daredevil behaviour would also be a means to gauge whether his actual skills at motorcycle riding are in line with his expectations, the accuracy of his distance and depth perception, and the reliability of his risk assessment based on perceptual judgements. As such, by making the motorcycle jump, Evel will be able to reduce uncertainty in various aspects within this novel environment.

Somewhat contrary to the FEP's self-preservation imperative, actually making the motorcycle jump—i.e., daredevil behaviour—is the most effective policy when it comes to testing the hypothesis of whether Evel can make a motorcycle jump across the canyon. Similarly, active inference holds that agents "resolve uncertainty through active sampling of the world" (Friston et al. 2017, 2639). It is of course possible to formulate safer speculative answers to the question of whether it is possible to leap across a canyon on a motorcycle. For instance, a physicist could make the necessary calculations, or a crash test lab could run simulations. However, they do not necessarily prove that Evel can make the jump on his motorcycle: they may leave out factors that are difficult to control for or that can go wrong, such as Evel faltering at some point, miscalculating his momentum and starting speed, or other forms of human error. Moreover, these solutions would lack a detailed interoceptive component for Evel, and thus may not be as convincing to him as going through with the jump. Thus, although informative, theoretical solutions would have less epistemic value than actually making the jump, which consequently proves to be the policy with the greatest epistemic value for the hypothesis Evel is trying to test and is therefore selected (Friston et al. 2017).

Dopamine deficit

The third solution to the daredevil challenge focuses on why daredevils persist in daredevil behaviour, and has as its starting point the effects of phenotypic diversity on free energy minimization (Friston 2010). According to the FEP canon, an organism's phenotype determines not only the repertoire of ecological conditions it can expect to occupy (Ramstead et al. 2019), but also the various regulatory processes necessary for it to maintain homeostasis (Kiverstein 2018). However, while conspecifics may have numerous traits in common, there is considerable variation at the individual level; consequently, what counts as surprising or stimulating differs from person to person, as will be illustrated below. One such variation can be found in the dopaminergic system, which may be caused by genetic (Nasiriavanaki et al. 2015) as well as developmental factors such as adolescence or experiences during infancy or early childhood (Steinberg 2008). Daredevil behaviour, especially of the third type (which is characterized by progressively risky behaviour) may be due to a deficiency in dopamine, which is "traditionally thought to report novelty, particularly in relation to action and expected value in the same setting" (Friston et al. 2014, 9).

Of particular prominence among the neurobiological features implicated in thrill-seeking and risk-taking in humans is the dopaminergic system, which plays a vital role in the regulation of affective and motivational processes (Sarshar et al. 2019; Steinberg 2008). It has been discovered that decreased functioning of the dopaminergic system can lead to an increase in sensation-seeking (Steinberg 2008), one of the

main components of thrill-seeking and risk-taking (Sarshar et al. 2019); this may be the case in daredevil behaviour. Diminished levels of dopamine have been linked to a decrease in the experience of typically rewarding stimuli as rewarding, often leading the individual to seek out novel or intense stimulation—such as those associated with street drugs or dangerous activities—that can provide a sufficient level of satisfaction (Steinberg 2008).

An individual with a dopaminergic system that does not produce enough levels of the neurotransmitter—i.e., a dopamine deficit—may be driven towards daredevil behaviour. This is due to needing stimulation more intense or novel than someone with normal dopamine levels in order to experience an activity as rewarding. In order to get their fix, the individual might then engage in daredevil behaviour, which is perceived to be much more thrilling than other activities that those with sufficient dopamine levels may find exciting enough. Once again let us take for instance Evel (who survived the motorcycle jump to figure in another example), a dopamine-deficient daredevil, and his friend Norm, who has a properly functioning dopaminergic system. Evel and Norm go to the amusement park and see a triple-loop roller coaster. While Norm is hesitant to get on the ride because he feels it is too scary, Evel cannot wait to try it out because he finds anything less terrifying bland and boring. Evel's increased dopamine level requirement thus influences selection of the action to go and ride the roller coaster with withs stomach churning-loops and steep drops.

After their initial roller coaster ride, Norm (who was scared out of his wits) has had enough of it for good, but Evel wants to relive the thrill. He thus decides to have another go…and yet another. Again, dopamine may have influenced this decision: it has been hypothesized that dopamine "could be in a position to promote (reinforce) the reselection (repetition) of recently selected actions or movements" (Redgrave and Gurney 2006, 971), particularly if the outcomes were "non-noxious (that is, novel or previously associated with reward" (Redgrave and Gurney 2006, 971). In this case, the roller coaster ride proved both novel and rewarding for Evel, fitting the aforesaid criteria for behaviour that is likely to be reselected.

After having ridden the roller coaster several times, Evel begins to find it boring. In fact, Evel realizes that none of his subsequent roller coaster rides were as exhilarating as the first one. He thus decides to up the ante of his daredevil behaviour, at first going on roller coasters with more loops or with steeper drops, moving on to riding them without the safety harness on (for now, let us pretend that this is possible). Evel's progressively risky behaviour is due to the fact that an activity that was once thrilling can eventually become dull through repetition as the daredevil habituates to the level of stimulation it affords. In order to achieve a comparable thrill, the daredevil may increase the riskiness of the action, since "[dopaminergic] responses to the predicted reward gradually diminish" (Redgrave and Gurney 2006, 967), where predicted reward pertains to the exhilaration caused by the action.

Upon trying the daredevil behaviour for the first time, the generative model encodes it as pleasurable, increasing the chances of its reselection. As such, a belief that the daredevil behaviour in question is enjoyable or thrilling is likewise encoded. Thus, when the daredevil action policy is reselected, a prediction is generated that it will result in exhilaration or other positive affective states. However, because the dopaminergic system has habituated, it fails to produce a response as intense as it

had in earlier instantiations of the daredevil behaviour in question. Consequently, prediction error or free energy to the effect of the daredevil behaviour failing to be as pleasurable as expected is generated, and must be quashed. Following the FEP's line of reasoning, there are two possibilities that may arise to quash the error signal. First, the generative model updates its beliefs and consequent predictions to now encode the daredevil behaviour as no longer exciting or enjoyable. Consequently, the daredevil action policy is less likely to be selected in the future, and the daredevil behaviour is abandoned (temporarily or permanently).

The second possibility—which accounts for persistent and intensifying daredevil behaviour—is that the generative model maintains its beliefs that that *type* of daredevil behaviour is exciting and enjoyable. Although this is inconsistent with the daredevil's actual experience that it is now less so, the relevant beliefs are not updated, and instead the daredevil takes measures to make the daredevil behaviour more stimulating. In other words, the daredevil engages in sensory-seeking behaviour. This can thus drive the daredevil to increase the riskiness of the behaviour, for instance by disregarding safety protocols. As such, while the type of daredevil behaviour remains fundamentally unchanged (e.g., driving too fast in the wrong direction, doing acrobatics on a balcony railing, swallowing copious amounts of desiccated cinnamon), certain details about it are altered to make the behaviour riskier and therefore more stimulating. The result is that subsequent instances (or *tokens*) of the type of daredevil behaviour concerned are more dangerous than previous ones.

Concluding remarks

The FEP places a heavy emphasis on homeostasis, optimization, surprisal avoidance, and self-preservation. However, contrary to its characterizations of biological systems, organisms do exist that exhibit actual behaviour that appear to seek out the very types of states that the FEP prescribes avoidance of. I have presented one such class of organisms: humans I refer to as daredevils. In order to account for their behaviour within a FEP framework, I have proffered three solutions, based on the current state of FEP literature. I intend these solutions to be conversation starters to be developed further in future FEP research, rather than finished products.

The purpose of this paper is to draw attention to daredevils in order to motivate the FEP to look more closely at organisms that do not, at first glance, appear to be in line with its explanatory framework. Doing so would enhance the FEP's ability to account for the diversity of behavioural characteristics of living organisms.

Acknowledgements My thanks go out to Alice Laciny, Ivan Gonzalez-Cabrera, Isabella Sarto-Jackson, Marco Treven, Emily C. Parke, Glenn Carruthers, Liz Schier, Ross Pain, Stephen Mann, and Michael David Kirchhoff. I am also grateful to the anonymous reviewers for their immensely helpful feedback. Work on this paper was divided between the Konrad Lorenz Institute for Evolution and Cognition Research (KLI) and the Zukunftskolleg at the University of Konstanz.

Funding Funding was received from the Konrad Lorenz Institute for Evolution and Cognition Research, and from the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments.

Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

Brodbeck J, Bachmann MS, Croudace TJ, Brown A (2012) Comparing growth trajectories of risk behaviors from late adolescence through young adulthood: An accelerated design. Dev Psychol 49(9):1732–1738

Caffray CM, Schneider SL (2000) Why do they do it? affective motivators in adolescents' decisions to participate in risk behaviours. Cogn Emot 14(4):543–576

Chitty D (1996) Do lemmings commit suicide?: Beautiful hypotheses and ugly facts. Oxford University Press, Inc., New York

Clark A (2013) Whatever next? predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36(3):181–253

Colombo M, Wright C (2018) First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*,

Friston K (2009) The free-energy principle: A rough guide to the brain? Trends Cogn Sci 13(7):293–301

Friston K (2010) The free-energy principle: A unified brain theory? Nat Rev Neurosci 11:127–138

Friston KJ, Lin M, Frith CD, Pezzulo G, Hobson JA, Ondobaka S (2017) Active inference, curiosity and insight. Neural Comput 29:2633–2683

Friston K, Kilner J, Harrison L (2006) A free energy principle for the brain. J Physiology-Paris 100:70–87

Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G (2015) Active inference and epistemic value. Cogn Neurosci 6(4):187–224

Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ (2014) The anatomy of choice: Dopamine and decision-making. Philosophical Trans Royal Soc B 369(1655):1–12

Friston K, Stephan KE (2007) Free-energy and the brain. Synthese 159(3):417–458

Friston K, Thornton C, Clark A (2012) Free-energy minimization and the dark-room problem. Front Psychol 3(130):1–7

Gu X, FitzGerald THB (2014) Interoceptive inference: Homeostasis and decision-making. Trends Cogn Sci 18(6):269–270

Hesp C, Constant A, Ramstead MJD, Badcock P (2019) A multi-scale view of the emergent complexity of life: A free-energy proposal. In: Georgiev GY, Smart JM, Martinez F, Claudio L, Price ME (eds) Evolution, development and complexity. Springer, Cham, pp 195–227

Hohwy J (2015) The neural organ explains the mind. In: Metzinger T, Windt JM (eds) Open MIND. MIND Group, Frankfurt am Main, pp 1–22

Kirchhoff M (2018) Predictive brains and embodied, enactive cognition: An introduction to the special issue. Synthese 195:2355–2366

Kiverstein J (2018) Free energy and the self: An ecological-enactive interpretation. Topoi 39:559–574

Laciny A, Zettel H, Kopchinskiy A, Pretzer C, Pal A, Salim KA et al (2018) Colobopsis explodens sp. n., model species for studies on "exploding ants" (hymenoptera, formicidae), with biological notes and first illustrations of males of the colobopsis cylindrica group. *ZooKeys*,. 751:1–40

Lelito JP, Brown WD (2006) Complicity or conflict over sexual cannibalism? male risk taking in the praying mantis tenodera aridifolia sinensis. Am Nat 168(2):263–269

Limanowski J, Blankenburg F (2013) Minimal self-models and the free energy principle. Front Hum Neurosci 7:1–12

Morehouse RE, Farley F, Youngquist JV (1990) Type T personality and the jungian classification system. J Pers Assess 54(12):231–235

Nasiriavanaki Z, ArianNik M, Abbassian A, Mahmoudi E, Roufigari N, Shahzadi S et al (2015) Prediction of individual differences in risky behavior in young adults via variations in local brain structure. Front NeuroSci 9(359):1–6

Prokop P, Václav R (2005) Males respond to the risk of sperm competition in the sexually cannibalistic praying mantis, mantis religiosa. Ethology 111(9):836–848

Rahnev D, Denison RN (2018) Suboptimality in perceptual decision making. Behav Brain Sci 41(223):1–66

Ramsden E, Wilson D (2014) The suicidal animal: Science and the nature of self-destruction. Past and Present 224(1):201–242

Ramstead MJD, Constant A, Badcock PB, Friston KJ (2019) Variational ecology and the physics of sentient systems. Phys Life Rev 31:188–205

Redgrave P, Gurney K (2006) The short-latency dopamine signal: A role in discovering novel actions? Nat Rev Neurosci 7(12):967–975

Sarshar M, Farley F, Fiorello CA, DuCette J(2019) T behavior: Psychological implications of thrill-seeking/risk-taking.Current Psychology,1–8

Schreiber A, Gimbel S (2010) Evolution and the second law of thermodynamics: Effectively communicating to non-technicians. Evolution: Educ Outreach 3:99–106

Self DR, Henry EDV, Findley CS, Reilly E (2007) Thrill seeking: The type T personality and extreme sports. Int J Sport Manage Mark 2(1–2):175–190

Seth AK (2013) Interoceptive inference, emotion, and the embodied self. Trends Cogn Sci 17(11):565–573

Seth AK (2014) Response to gu and FitzGerald: Interoceptive inference: From decision-making to organism integrity. Trends Cogn Sci 18(6):270–271

Steinberg L (2008) A social neuroscience perspective on adolescent risk-taking. Dev Rev 28(1):78–106

The Editors of Encyclopaedia Britannica (2020) *Evel knievel.* Retrieved 18/11, 2020, from https://www.britannica.com/biography/Evel-Knievel

Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: The organization of living systems, its characterization and a model. BioSytems 5(4):187–196