Risks Deriving from the Agential Profiles of Modern AI Systems

Barnaby Crook, University Bayreuth, barnaby.crook@uni-bayreuth.de

Abstract

Modern AI systems based on deep learning are neither traditional tools nor full-blown agents. Rather, they are characterised by idiosyncratic *agential profiles*, i.e., combinations of agency-relevant properties. Modern AI systems lack superficial features which enable people to *recognise* agents but possess sophisticated information processing capabilities which can undermine human goals. I argue that systems fitting this description, when they are adversarial with respect to human users, pose particular risks to those users. To explicate my argument, I provide conditions under which agential profiles are explanatorily relevant to harms caused. I then contend that the role of recommender systems in producing harmful outcomes like digital addiction satisfies these conditions.

> *"While [humans] are very good at recognizing agency in both the three-dimensional world of conventional behavior and the much higher dimensional space of social interactions, we are poor at recognizing intelligence in novel guises" (Fields & Levin, 2022, p. 2)*

Introduction

Modern artificial intelligence (AI) systems based on deep learning process large volumes of data and learn complex representations supporting adaptive, goal-directed behaviours (LeCun et al., 2015; Rahwan et al., 2019). Such systems are, in virtue of these properties, markedly more agential than traditional tools (Russell & Norvig, 2020). At the same time, however, modern AI systems lack core aspects of biological agency such as embodiment and autonomy, sharply distinguishing them from living organisms (Meincke, 2018; Moreno & Etxeberria, 2005). This state of affairs has prompted renewed philosophical discussion over the applicability of agency to AI systems (e.g., Nyholm, 2018; Swanepoel, 2021). Rather than litigating the applicability question, I argue in this paper that the distinctive combination of agency-relevant properties possessed by modern AI systems, which I describe as an *agential profile*, plays an important and underappreciated role in their potential to cause harms to human users. In particular, I observe that modern AI systems possess qualitatively novel agential profiles, combining a paucity of the superficial features of agency with sophisticated goal-directed information processing capabilities. Such profiles, I claim, thwart inferential reasoning

via extant concepts and are thus apt to subvert intuitive human judgements and undermine higher-order human goals. Consequently, appreciating agential profiles is necessary to understand and prevent harms as AI systems continue to proliferate.

The structure of this paper is as follows: In section one, I introduce the idea of agential profiles and specify a profile for a prototypical modern AI system.[1] In section two, I argue that this agential profile generates novel risks for human users. Specifically, the dissociation of features which enable *recognition* of adversarial agency and those which demand mindful *negotiation* of it produces scenarios in which users interact with adversarial systems unknowingly, increasing the risk of having their goals undermined. In section three, I discuss digital addiction as a specific harm for which, I claim, agential profiles play an explanatory role. In section four, I clarify the scope of my claims, counter possible objections, and make tentative proposals for ameliorative action. I then conclude.

Section 1: Agential Profiles

In this section I introduce *agential profiles*. Given a set of *dimensions* of agency, viz., agency-relevant ways in which systems can vary, the agential profile of any system is a specification of the extent to which it exhibits each of these agency-relevant properties.

1.1 Dimensions of Agency

Agency is a conceptual term invoked in disciplines as diverse as philosophy (Schlosser, 2019; Swanepoel, 2021), sociology (Nikolic & Kasmire, 2013; Winner, 1977), biology (Monod, 1971; Okasha, 2018), psychology (Bandura, 2006; Carey, 2009), and artificial intelligence (Meincke, 2018; Moreno & Etxeberria, 2005; Russell & Norvig, 2020). Analyses of agency vary enormously with respect to which properties are deemed constitutive of it and the degree to which they must be present for a system to count as agential. In this paper, I adopt a *dimensional* approach to agency (c.f., Dung, 2024; Okasha, 2018). Such an approach is desirable because it is informationally rich (compared to categorical approaches which treat agency as a discrete property) and reflects the heterogeneity of the agency concept. Accordingly, I begin by collating properties of agency proposed across disciplines. Since thorough coverage of all disciplines would be impossible, this is a selective survey. However, by drawing on prominent and diverse accounts, I aim for a representative selection of scholarly views.

Philosophers, motivated by providing accounts of agency that ground actions and moral responsibility, have proposed numerous dimensions. These include *initiative* (Schlosser, 2019), *intentionality* and *beliefs* (Davidson, 2001), *plans* (Bratman, 2000),

[1] I use the term *modern AI system* throughout the text. When I use this term, I have in mind systems trained on large datasets via machine learning. In principle, however, the argument applies to any system with the right kind of agential profile (see Section 1), regardless of the methodology used to develop it.

normativity and *internal coherence* (Korsgaard, 2008), *reflective evaluation* (Frankfurt, 1971), *unity-of-purpose* (Kennett & Matthews, 2003), and *consciousness* (Swanepoel, 2021). Biologists, aiming to understand the evolutionary roots of agency and use it to explain organismic behaviour, focus on aspects of agency including *behavioural flexibility* (Monod, 1971), *adaptedness* (Okasha, 2018), *embodiment* (Cisek, 2019), *self-maintenance* (Maturana & Varela, 1980), *autonomy* (Moreno & Etxeberria, 2005), and *persistence-over-time* (Godfrey-Smith, 2009). Social scientists have applied the concept of agency to explain the behaviour of sociotechnical systems. To do so, they invoke properties such as *goals* (Perrow, 1991), *plans* (Simon, 1979), *hierarchical organisation* (Winner, 1977), and *centralised control* (Nikolic & Kasmire, 2013). Psychologists have worked on both the felt sense of agency and the attribution of agency to other entities. Properties relevant for the former include *volition* (Jeannerod, 2006), *self-efficacy* (Bandura, 2006) and *control* (Skinner, 1996) while those relevant for the latter include *goal-directedness* (Heider & Simmel, 1944), *embodiment* (Carey & Spelke, 1994), *animacy* (Gergely & Jacob, 2012), and *spatiotemporal continuity* (Carey, 2009). Finally, AI researchers have attempted to develop artificial agents. Dimensions foregrounded in this pursuit include *perception* (Russell & Norvig, 2020), *action* (Brooks, 1991), *experience-dependent behaviour* (Wang, 2019), *learning* (Sutton & Barto, 2020), *sensorimotor coupling* (Beer, 1995), and *adaptivity* (Rosenblueth et al., 1943).

Many properties invoked as agency-relevant across disciplines either recur or exhibit conceptual overlap. Thus, through combining and restructuring terms, the dimensionality can be reduced to a manageable level. To do this, I attempted to find a minimal set of dimensions which capture variation in the systems discussed in the literature. Though far from perfect, I suggest the following dimensions suffice to capture most relevant variation (related terms in parentheses) (see Dung, 2024 for an alternative selection):

Autonomy: Initiating actions without external elicitation (volition, control, initiative, action, self-efficacy, animacy).

Goal-directedness: Behaving so as to reliably bring about particular states of affairs across varied contexts (goals, unity-of-purpose, intentionality).

***Reflexivity*:** Reflecting on preferences, goals, and reasons (reflective evaluation, internal coherence, beliefs, consciousness).

Structural Coherence*:* Being organised such that distinct parts cooperate with one another (hierarchical organisation, centralised control, unity-of-purpose, internal coherence).

Embodiment*:* Being constituted by a physical body embedded in an environment (spatiotemporal continuity, self-maintenance, perception, sensorimotor coupling, autonomy).

Flexibility: Exhibiting context-dependent behaviour (adaptedness, sensorimotor coupling).

***Learning*:** Exhibiting experience-dependent behaviours (adaptivity, normativity, centralised control).

Temporal Coherence: Behaving such that earlier and later actions are coordinated with each other (planning, persistence-over-time, spatiotemporal continuity).

I take these dimensions to constitute a reasonable basis for characterising agential profiles.

1.2 Agential Profiles of Modern AI Systems

Let us now assess the credentials of deep learning-based AI systems with respect to the dimensions of agency adduced above. For concreteness, consider a deep neural network trained with reinforcement learning to control video recommendations a user sees on a platform like YouTube (e.g., M. Chen et al., 2019; X. Chen et al., 2023).[2] The system is trained to maximise *reward*, a numerical value here equated with user watch time (sometimes euphemistically described as *user satisfaction*).[3] Since my aim is to characterise the agential profile of such a system roughly, I will not commit to specific views on how each dimension should be understood and quantified. Rather, I will provide general considerations, allowing room for theoretical uncertainty.

Our recommender system scores highly on *goal-directedness*. Since the system has been trained to maximise a specific, quantifiable objective function, its actions are *all* tailored to achieving that objective.[4] Similarly, our system is strong in *structural coherence*. Deep neural networks are often trained end-to-end (LeCun et al., 2015), meaning their entire internal structure is optimised simultaneously with respect to a single objective. This training scheme ensures that the parts of our system (e.g., hierarchically organised nodes and layers of the neural network) operate in a highly coordinated fashion. For *learning***,** our system also does fairly well. Recommending suitable videos to users is a difficult problem due to the sparsity of the data (most users never encounter most videos) and the enormity of the action space (M. Chen et al., 2019). Our system effectively handles large state spaces, models complex user preferences, and learns prospective strategies to optimise long-term user engagement (Evans & Kasirzadeh, 2023; Franklin et al., 2022). However, our system is less sample-

[2] See Section 4 for elaboration on this example.
[3] More precisely, because it evaluates sequences of actions and (predicted) user responses, the system is trained to maximise expected *cumulative* reward, for some specific temporal horizon (with a discount rate).
[4] Though note that the mathematical formulation of the objective may depart from the intended objective of the system's developers.

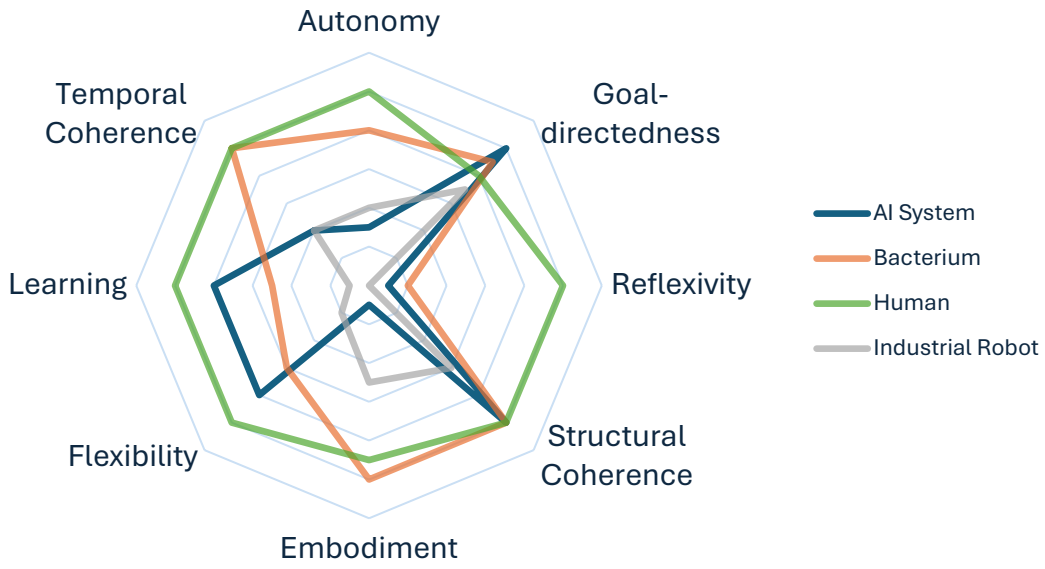
efficient than biological learners and cannot generalise to distinct tasks (Zador, 2019). Likewise, when it comes to *flexibility*, our system scores moderate to high. In one sense, the system is highly flexible. It operates over an enormous, ever-changing action space (i.e., recommendable items) and tailors its behaviour precisely to representations of user states (M. Chen et al., 2019). Further, our recommender system is highly adapted, pursuing its objective effectively enough to persist (viz., remain deployed, c.f., Rahwan et al., 2019). However, unlike humans and other organisms, our system cannot actively expand the *types* of actions it can perform through exploration and evolution (Roli et al., 2022).

For other dimensions, our recommender system does much less well. It scores low to moderate on *autonomy*. On the positive side, it operates without human intervention once deployed, selecting and recommending items without supervision. However, the system is fundamentally *reactive*, acting only in response to user prompting (Okasha, 2018). Since initiating actions is central to autonomy, this deficiency is significant. Similarly, for *temporal coherence* our system receives a low to moderate score. It is able to generate, evaluate, and implement *trajectories* of actions (X. Chen et al., 2023; Evans & Kasirzadeh, 2023), a key component of short-term temporal coherence. However, our system has no episodic, working, or long-term memory with which to ground persistence-over-time, i.e., identity (Jablonka & Ginsburg, 2022). Further, because its actions are reactive, our system remains dormant unless prompted, ruling out organism-like temporal coherence. When it comes to *reflexivity,* our system scores very poorly. Reflexive processing is a demanding notion involving active contemplation of one's goals and beliefs and, plausibly, metacognitive mechanisms related to consciousness (Dehaene et al., 2017). While a computational implementation of reflexivity may be possible, no such machinery is present in current-generation recommender systems. Finally, our system also scores very poorly in *embodiment*. To be sure, a recommender system must be implemented in some kind of physical structure. However, our system does not possess a manipulable body, does not navigate a structured spatial environment (pace Fields & Levin, 2022), is not spatiotemporally continuous, and does not maintain the physical conditions for its continued existence (Roli et al., 2022).

To illustrate, Figure 1 contrasts the agential profile of a modern AI system with a bacterium, a human being, and an industrial robot.[5]

[5] This chart is purely illustrative. The values chosen to specify each system's profile are rough estimates and the entities graphed obviously exhibit considerable variation (e.g., different bacterial species and industrial robots will, in reality, have different profiles).

Figure 1: Agential Profiles Compared



Notice that our AI system has an *idiosyncratic* agential profile. That is, it differs qualitatively (in shape) from other kinds of entities. While Figure 1 only depicts some choice contrasts, I contend that, were we to plot more entities – further biological organisms, partially autonomous tools, and the like – the profile of modern AI systems would remain truly distinctive, exhibiting limited overlap with any other system.

Section 2: Risks from the Agential Profiles of Modern AI Systems

In this section I argue that the agential profiles of modern AI systems lead to distinctive risks for human interactants. I start by distinguishing between dimensions of agency that are important for *recognising* other agents and dimensions that are important for our ability to *negotiate* agency. I then note that modern AI systems can be *adversarial* with respect to human users in the sense that they pursue incompatible goals in a shared environment. Finally, I outline a set of conditions which, when satisfied, render agential profiles explanatorily relevant to harmful outcomes.

2.1 Recognising and Negotiating Agency

As we have seen, agency is a multi-dimensional concept. One way in which its constitutive dimensions differ is in the role they play in inducing *recognition* of agency. Scholars across disciplines have pointed out that human beings attribute agency to external entities promiscuously (Barrett, 2004; Carey & Spelke, 1994; Fields, 2014;

Heider & Simmel, 1944). Crucially, research from developmental psychology and neuroscience posits agency recognition as *implicit*, occurring automatically as part of the unconscious cognitive processing of sensory information (Fields, 2014). On this view, agency is an ontogenetically primitive, pre-linguistic conceptual category supporting inferential reasoning about external objects (Seyfarth & Cheney, 2013; Spelke, 2022). When I speak of recognising agency, it is this implicit sense I intend. Note that while agential *profiles* are multidimensional, the *recognition* of agency, as a cognitive category judgement, may still be binary or unidimensional.

Research into which properties trigger recognition of agency suggests important roles for embodiment, autonomy, goal-directedness, and temporal coherence (Carey, 2009; Fields, 2014; Gergely & Jacob, 2012). Generally, some combination of these properties is required for agency to be attributed (Carey & Spelke, 1994). With respect to embodiment, this may also include entities embodied within virtual environments. For autonomy, the contribution of animacy, in the sense of self-propelled motion, is crucial (Spelke, 2022). Goal-directedness, viz., perceiving that some entity reliably brings about particular states of affairs, is another important dimension.[6] Finally, the temporal coherence of an entity plays a crucial role in allowing an observer to ascribe it *any* stable properties, including that agency (Fields, 2014).

For some dimensions of agency, a system's scoring highly on those dimensions makes it challenging for another entity to *negotiate* the agency of that system. I use the term *negotiate* to mean continuing to achieve one's goals when interacting with that system. For example, if I encounter a stray dog while running, I may proceed cautiously to ensure that I can pass safely. If the dog's behaviour (e.g., growling) indicates that it prefers me not to pass, I may alter my plan rather than risk being harmed. Most of the properties we exploit to recognise agential systems are also relevant to the challenge of negotiating the agency of those systems. A system embodied in the real-world can impinge upon goals that disembodied entities are unlikely to influence (e.g., my run). Similarly, autonomous systems pose a greater threat than reactive ones because they can act unprompted. However, crucially, part of what makes certain agential systems challenging to negotiate comes from dimensions which are *not* superficially observable. Human agency is challenging to negotiate not only because people are embodied, autonomous, and temporally coherent, but also because they learn quickly, behave flexibly, and effectively pursue goals. Modern AI systems may not be embodied or autonomous, but their flexibility, learning, structural coherence, and goal-directedness can still render their agency challenging to negotiate.

If my analysis is right, the dimensions of agency most important for our recognition of agents and those which make the negotiation of agency challenging are, at least

[6] Indeed, some degree of goal-directedness is a plausible candidate for a necessary condition on agency.

partially, dissociable. In order to show how this dissociation generates novel risks, we need to introduce the notion of adversariality.

2.2 When are AI Systems Adversarial?

Adversariality is a two-place predicate. That is, it cannot be attributed to an isolated system, but only to one system with respect to another. I employ the term functionally, meaning its attribution depends only on the behavioural impact of interactions, not on details about *how* effects are brought about or the intent with which a system was developed. Clearly, two systems are adversaries if they directly compete, as in games like chess (Russell & Norvig, 2020). Intuitively, this is because one system achieving its goal precludes the other system from doing so. However, a more general definition ought also to capture *partially competitive* cases in which System A achieving states (internal or external) that are high in its preference ordering reduces the probability of System B doing so, such as aggressive resource accumulation in an environment with limited supply. Clearly, such a definition applies to interactions between systems that explicitly estimate and represent the value of states, as chess-playing systems do with respect to board positions. However, we wish to apply the definition to people, for whom the relevant notion of preferences over states is less clearly defined. Thus, more must be said about what it means for a person to achieve preferable states.

For the purposes of my argument, the key notion is that of *higher-order goals*. Higher-order goals are consciously chosen (and periodically re-evaluated) during deliberate reflection and their satisfaction requires coordinating actions across time. Such goals often come in the form of values (e.g., *health, career, relationships*) and imply normative standards against which local outcomes are assessed. For example, if Alice considers whether her day went well, her answer will depend upon the extent to which her behaviours contributed to the satisfaction of her higher-order goals. Notably, some such theoretical posit is *required* to make sense of the idea that people can behave contrary to their own best interests (where those interests are identified with higher-order goals) (Franklin et al., 2022; Sen, 1977). The alternative, revealed preference theory, renders all uncoerced behaviour tautologically preference-satisfying.

With the idea of higher-order goals in mind, we can return to adversariality. I define it thus (modified from Russell & Norvig, 2020, p. 111):

Adversariality: System A is adversarial with respect to System B if A behaves as if maximising a performance measure whose value depends on B's behaviour being contrary to B's higher-order goals.

Note that the use of the qualifier *as if* ensures that this definition depends only on behaviour, i.e., A need not represent B's higher-order goals in order for A to be

adversarial with respect to B. Further, to a first approximation, B's behaviour, x, is *contrary* to its higher-order goal, y, if the probability of B achieving y, given they x-ed, is lower than the probability of B achieving y, had they x'-ed, where x' is a reasonable alternative to x.

2.3 Risks from Agential Profiles: A Mechanism

With the relevant conceptual machinery in place, I now explain how AI systems' agential profiles pose risks to human interactants. I suggested above that we rely on particular features to *recognise* agency in other systems (Carey, 2009; Gergely & Jacob, 2012). Returning to Figure 1, we see that modern AI systems do not score highly on these dimensions. Thus, we should not expect humans interacting with systems fitting this agential profile to recognise them as agents.[7] We have also seen that modern AI systems score highly on features that make it challenging to *negotiate* agency. These features are particularly crucial for enabling efficacious goal-directed behaviour. Therefore, we should expect systems fitting this agential profile to be effective at achieving their preferred states. When such systems are also *adversarial* with respect to human interactants, this engenders risk. The dissociation of features that enable *recognition* of agency and those that demand *negotiation* of agency brings about scenarios in which people engage in interaction with adversarial agential systems unknowingly. In such cases, especially with highly capable AI systems, there is a danger of human interactants' higher-order goals being subverted.

To spell out the idea further, notice that *recognising* agency is instrumentally valuable for *negotiating* it. To sketch a mechanism, when we recognise the presence of (potentially adversarial) agency, we deploy cognitive resources to considering how our own goals might be undermined and take actions to ensure they are not (either through avoidance, aggressive actions, or compromise). When we do not recognise the presence of agency, we are less inclined to worry that we may have our higher-order goals undermined. This lack of caution increases the probability of harms, such as the formation of maladaptive behavioural habits (viz., those which do not contribute to the achievement of higher-order goals). This analysis is in line with Bayer and colleagues' observation that habits form and endure due to "reduced self-surveillance over one's behaviour" (2022, p. 3). Naturally, a more detailed mechanism is required to explain specific cases, but this simple account captures the core dynamic.

To be precise, I argue that pursing one's (higher-order) goals in the presence of an adversarial system which scores highly on the negotiation dimensions of agency will be

[7] Here, and in what follows, I treat recognition of agency as binary. A graded view of recognising agency can be accommodated without changing the structure of the argument. To do so, swap out "recognition of agency" for "recognition of agency to the degree required to negotiate it".

more successful, ceteris paribus, when we recognise the system as an agent than when we do not, even if that system does not score highly on the recognition dimensions of agency. This is why I contend that it is the agential profile *as a whole* that poses a specific and novel risk for human users. Perhaps counterintuitively, that modern AI systems *lack* the properties that ground the human ability to recognise agency contributes to their potential to undermine human goals. To make the claim clearer, I propose four conditions which, if satisfied, render the agential profile of an AI system explanatorily relevant to harms suffered by a human subject:

(1) The subject must experience harms as a result of behaving contrary to their higher-order goals in an interaction with an AI system.
(2) The AI system must behave as though it is maximising a performance measure whose value depends on the subject's behaviour being contrary to her higher-order goals.
(3) The harms must be dependent on the system's (positive) agential properties. That is, had the AI system lacked its agential properties, the harms would not have occurred (as severely).
(4) The harms must be dependent on the subject's failure to recognise the AI system as agential (to a suitable degree). That is, had the subject recognised the AI system as agential, the harms would not have occurred (as severely).

Condition 1 ensures there are harms to explain. Condition 2 introduces adversariality. Condition 3 is a counterfactual ensuring that the agential properties of the AI system are relevant to causing the harms.[8] And condition 4 is a counterfactual that must hold for the agential profile *as a whole* to be explanatorily relevant (rather than just the subset of positive, negotiation-relevant agential properties).

Section 3: Digital Addiction as a Case Study

The argument above was presented abstractly. A concrete example will help demonstrate that the risks are plausible. In particular, I suggest that conditions 1-4 are satisfied by cases of habitual overuse of digital technologies involving AI recommendation systems (e.g., Hasan et al., 2018). A growing body of research shows increasing rates of addiction to digital technologies (Cerniglia et al., 2017; Meng et al., 2022). These outcomes are partially driven by the use of AI systems designed to maximise user time spent (Bayer et al., 2022; Chianella, 2021). To assess whether it is

[8] This rules out ascribing explanatory roles to the agential profiles of simpler systems which cause harms due to malfunction or misuse.

plausible that agential profiles play an explanatory role, let us walk through the conditions laid out above.

First, are human users experiencing harms as result of behaving contrary to their higher-order goals when interacting with AI systems? Qualitative and quantitative evidence strongly suggests so. Many users report spending more time than they would like interacting with digital platforms which use AI systems to organise content (Tokunaga, 2017). As Bayer and colleagues put it, such "habits become problematic when specific habit sequences consistently undercut users' goals" (2022, p. 7). Further, excessive use of digital platforms supported by recommender systems is correlated with negative mental health outcomes (Cai et al., 2023). To relate this to the discussion of adversariality in section 2.2, it seems clear that excessive use of digital technologies can be contrary to users' higher-order goals in the relevant sense (with refraining from use constituting a *reasonable alternative* to the goal-undermining behaviour).

Second, do the AI systems in question behave as if maximising a performance measure dependent on the behaviour of human users being contrary to their higher-order goals? Again, this condition is likely satisfied. The advertisement-based economic model on which many digital media platforms operate depends upon capturing user attention (Davenport & Beck, 2001). And technical research demonstrates the efficacy with which AI systems leverage large volumes of user data to deliver absorbing content (M. Chen et al., 2019; Evans & Kasirzadeh, 2023). Thus, although the specific performance measures used to train AI systems actually operating on digital media platforms are often proprietary, it is clear that they behave as though maximising a measure correlated with (if not identical to) user time spent. Naturally, this only satisfies the condition given excessive usage is contrary to users' higher-order goals. However, this is addressed by condition one.

Third, are the harms dependent on the AI systems' positive agential properties? Evidence from empirical studies suggests so. For example, Hasan and colleagues found that "use of recommender systems has a significant positive influence on excessive usage of video websites" (2018, p. 226). And Cao and colleagues found that "personalization positively influence[s] individuals' emotional and functional attachment on social media, thereby causing addictive behavior (2020, p. 1320)". Recommender systems' positive agential properties, viz., goal-directedness, flexibility, and learned strategies, are the most plausible explanation for these effects.

Finally, are the harms dependent on human users' failure to recognise the agency of the system they are interacting with? This is the most contentious condition as what recognising a system as agential means is conceptually subtle and difficult to measure. One relevant line of research comes from studies which manipulate perceptions of system behaviours by varying how the systems are described. If (mis)perceptions of system agency can influence user behaviour, we should expect different system

descriptions (which vary in their association with agency) to produce an effect. This is indeed what researchers have found (Candrian & Scherer, 2024; Langer et al., 2022). However, this is merely suggestive evidence. Further empirical work is required to rigorously evaluate whether (and in which range of cases) condition four is satisfied.

If my analysis is right, recommender systems' role in digital addiction plausibly exemplifies the distinctive risks posed by the agential profiles of modern AI systems.

Section 4. Discussion

Several points require further discussion. First, I have focused here on risks stemming from *underestimating* the agency of AI systems. Other scholars have worried about the opposite, viz., risks stemming from *overestimating* the agency of such systems (e.g., Placani, 2024). These positions are not incompatible. My view is that *no* familiar conceptual hook (e.g., *tool* or *agent*) induces effective intuitive reasoning about current AI systems. Accordingly, both of these risks can be considered as special cases of *misperceiving* the agency of AI systems, liable to occur in different contexts. Relatedly, a reviewer worries that framing the concern in terms of *underestimation* is contrary to the dimensional view of agency articulated by agential profiles. Here, I stress again the distinction between the *recognition* of agency, which is a subconscious cognitive judgement (that may be binary or graded on a single dimension), and the agency-relevant properties systems in fact possess, which require many dimensions to express. Thus, while I agree that a nuanced, multidimensional view of AI systems is beneficial in philosophical and scientific contexts, my argument concerns how these profiles interact with intuitive cognitive judgements (which is where *underestimation* comes in).

Second, the mechanism sketched in section 2.3 is general and may apply to numerous present and future human-computer interaction scenarios. As such, though digital addiction is an important issue, my argument is not for the narrower claim that such a mechanism explains that harm in particular, but for the broader claim that novel agential profiles pose distinctive risks. Indeed, given the proposed mechanism routes through a mismatch between our intuitive reasoning faculties and the properties of AI systems, we should not be surprised to find difficult-to-identify scenarios (either extant or future) which, upon close inspection, turn out to fit the pattern.

Third, my argument in this paper is that modern AI systems' agential profiles help explain why they are liable to subvert human goals. This does not mean that agential profiles are *sufficient* to explain harms caused. In line with a pluralist approach to explanation, I view the mechanism I have described in this paper as complementary to explanations of maladaptive habit formation in terms of brain mechanisms (e.g.,

Serenko & Turel, 2022), social and psychological factors (e.g., Bayer et al., 2022), and further aspects of the design of digital technologies (e.g., Chianella, 2021).

Next, I address possible objections. Though my argument contains several contestable steps, I believe the conclusions are robust to disagreement on many details. For example, one may disagree with the dimensions I chose to capture agential profiles. However, my argument does not depend upon those specific dimensions. Other choices, I posit, would also reveal the (partial) dissociation of recognition and negotiation dimensions in modern AI systems. Similarly, one might disagree with my claims about the recognition of agency and its role in cognitive behaviour, perhaps denying that the category is innate (see Buckner, 2023 for discussion). However, this is compatible with the claim that the implicit recognition of agents plays an important role in the human cognitive system, even if that capacity is learned during development.

Further, though I presented theoretical arguments to support my claims, I acknowledge that more empirical work is needed to assess whether and to what degree they hold in practice. In particular, studies should further investigate how varying the dimensions of agency that are critical for recognition affects user behaviour. Per my argument, inducing the recognition of agency ought to yield more cautious and mindful behaviour.

Finally, if my claims are borne out empirically, there are implications for ameliorative policies. For example, clear cues indicating that one is interacting with an agential system may induce more prudential behaviour, reducing the risk of maladaptive habit formation. Should this be demonstrated, legislation could compel companies deploying AI in user-facing applications to notify users that they are interacting with (partially) agential systems. In the longer term, numerous AI systems with idiosyncratic and perplexing agential profiles seem likely to be developed. Assuming this is so, continual refinement of our conceptual understanding will be required. Moving beyond the dichotomy of tools and agents may be necessary to protect users from further risks. To this end, agential profiles can serve as a valuable framework capturing a nuanced picture of AI systems and making sense of the risks they pose.

5. Conclusion

In this paper I used the conceptual apparatus of *agential profiles* to argue that modern AI systems pose particular risks to human users. I argued that disembodied AI systems based on deep learning dissociate dimensions of agency that have, historically, co-occurred. In particular, these systems lack the properties that we rely on to recognise agency while possessing those that can undermine our ability to negotiate that agency. When AI systems are adversarial with respect to their human users, this scenario threatens to cause harms. I supported my argument with the case study of digital

addiction and AI-based recommender systems. I claimed that the agential profiles of such systems play an underappreciated role in explaining how and why they induce harmful outcomes like digital addiction. If my argument is correct, intuitive human reasoning about agency is ill-suited to cope with modern AI systems. Active steps must be taken to ensure we can negotiate their distinctive agential profiles.

References

Bandura, A. (2006). Toward a Psychology of Human Agency. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *1*(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Barrett, J. L. (2004). *Why would anyone believe in God?* AltaMira Press.

Bayer, J. B., Anderson, I. A., & Tokunaga, R. S. (2022). Building and breaking social media habits. *Current Opinion in Psychology*, *45*, 101303. https://doi.org/10.1016/j.copsyc.2022.101303

Beer, R. D. (1995). A Dynamical Systems Perspective on Agent-Environment Interaction. *Artificial Intelligence*, *72*(1–2), 173–215. https://doi.org/10.1016/0004-3702(94)00005-l

Bratman, M. E. (2000). Reflection, Planning, and Temporally Extended Agency. *The Philosophical Review*, *109*(1), 35–61. https://doi.org/10.2307/2693554

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Buckner, C. J. (2023). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Cai, Z., Mao, P., Wang, Z., Wang, D., He, J., & Fan, X. (2023). Associations Between Problematic Internet Use and Mental Health Outcomes of Students: A Meta-analytic Review. *Adolescent Research Review*, *8*(1), 45–62. https://doi.org/10.1007/s40894-022-00201-9

Candrian, C., & Scherer, A. (2024). How Terminology Affects Users' Responses to System Failures. *Human Factors*, *66*(8), 2082–2103. https://doi.org/10.1177/00187208231202572

Cao, X., Gong, M., Yu, L., & Dai, B. (2020). Exploring the mechanism of social media addiction: An empirical study from WeChat users. *Internet Research*, *30*(4), 1305–1328. https://doi.org/10.1108/INTR-08-2019-0347

Carey, S. (2009). Core Cognition: Agency. In S. Carey (Ed.), *The Origin of Concepts* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367638.003.0005

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In *Mapping the mind: Domain specificity in cognition and culture* (pp. 169–200). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.008

Cerniglia, L., Zoratto, F., Cimino, S., Laviola, G., Ammaniti, M., & Adriani, W. (2017). Internet Addiction in adolescence: Neurobiological, psychosocial and clinical issues. *Neuroscience and Biobehavioral Reviews*, *76*(Pt A), 174–184. https://doi.org/10.1016/j.neubiorev.2016.12.024

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., & Chi, E. H. (2019). Top-K Off-Policy Correction for a REINFORCE Recommender System. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 456–464. https://doi.org/10.1145/3289600.3290999

Chen, X., Yao, L., McAuley, J., Zhou, G., & Wang, X. (2023). Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems*, *264*, 110335. https://doi.org/10.1016/j.knosys.2023.110335

Chianella, R. (2021). Addictive digital experiences: The influence of artificial intelligence and more-than-human design. *Blucher Design Proceedings*, *9*(5), 414–425. https://www.proceedings.blucher.com.br/article-details/36959

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265–2287. https://doi.org/10.3758/s13414-019-01760-1

Davenport, T. H., & Beck, J. C. (2001). The Attention economy. *Ubiquity*, *2001*(May), 1. https://doi.org/10.1145/376625.376626

Davidson, D. (2001). *Essays on actions and events* (2nd ed). Clarendon Press ; Oxford University Press.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dung, L. (2024). Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010. https://doi.org/10.1093/pq/pqae010

Evans, C., & Kasirzadeh, A. (2023). User Tampering in Reinforcement Learning Recommender Systems. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 58–69. https://doi.org/10.1145/3600211.3604669

Fields, C. (2014). Motion, identity and the bias toward agency. *Frontiers in Human Neuroscience*, *8*, 597. https://doi.org/10.3389/fnhum.2014.00597

Fields, C., & Levin, M. (2022). Competency in Navigating Arbitrary Spaces as an Invariant for Analyzing Cognition in Diverse Embodiments. *Entropy*, *24*(6), Article 6. https://doi.org/10.3390/e24060819

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Franklin, M., Ashton, H., Gorman, R., & Armstrong, S. (2022). *Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI* (arXiv:2203.10525). arXiv. https://doi.org/10.48550/arXiv.2203.10525

Gergely, G., & Jacob, P. (2012). Reasoning about instrumental and communicative agency in human infancy. *Advances in Child Development and Behavior*, *43*, 59–94. https://doi.org/10.1016/b978-0-12-397919-3.00003-4

Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection*. Oxford University Press.

Hasan, M. R., Jha, A. K., & Liu, Y. (2018). Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Computers in Human Behavior*, *80*, 220–228. https://doi.org/10.1016/j.chb.2017.11.020

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Jablonka, E., & Ginsburg, S. (2022). Learning and the Evolution of Conscious Agents. *Biosemiotics*, *15*(3), 401–437. https://doi.org/10.1007/s12304-022-09501-y

Jeannerod, M. (2006). From Volition to Agency: The Mechanism of Action Recognition and Its Failures. In N. Sebanz & W. Prinz (Eds.), *Disorders of Volition* (pp. 175–192). The MIT Press. https://doi.org/10.7551/mitpress/2457.003.0010

Kennett, J., & Matthews, S. (2003). The Unity and Disunity of Agency. *Philosophy, Psychiatry, & Psychology*, *10*(4), 305–312.

Korsgaard, C. M. (2008). *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press.

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28. https://doi.org/10.1145/3491102.3517527

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living* (Vol. 42). Springer Netherlands. https://doi.org/10.1007/978-94-009-8947-4

Meincke, A. S. (2018). Bio-Agency and the Possibility of Artificial Agents. In A. Christian, D. Hommen, N. Retzlaff, & G. Schurz (Eds.), *Philosophy of Science: Between the Natural Sciences, the Social Sciences, and the Humanities* (pp. 65–93). Springer International Publishing. https://doi.org/10.1007/978-3-319-72577-2_5

Meng, S.-Q., Cheng, J.-L., Li, Y.-Y., Yang, X.-Q., Zheng, J.-W., Chang, X.-W., Shi, Y., Chen, Y., Lu, L., Sun, Y., Bao, Y.-P., & Shi, J. (2022). Global prevalence of digital addiction in general population: A systematic review and meta-analysis. *Clinical Psychology Review*, *92*, 102128. https://doi.org/10.1016/j.cpr.2022.102128

Monod, J. (1971). *Chance and necessity: An essay on the natural philosophy of modern biology* (1st American ed.). Knopf.

Moreno, A., & Etxeberria, A. (2005). Agency in natural and artificial systems. *Artificial Life*, *11*(1–2), 161–175. https://doi.org/10.1162/1064546053278919

Nikolic, I., & Kasmire, J. (2013). Theory. In K. H. van Dam, I. Nikolic, & Z. Lukszo (Eds.), *Agent-Based Modelling of Socio-Technical Systems* (pp. 11–71). Springer Netherlands. https://doi.org/10.1007/978-94-007-4933-7_2

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Okasha, S. (2018). *Agents and Goals in Evolution*. Oxford University Press. https://doi.org/10.1093/oso/9780198815082.001.0001

Perrow, C. (1991). A Society of Organizations. *Theory and Society*, *20*(6), 725–762.

Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00419-4

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy', … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), Article 7753. https://doi.org/10.1038/s41586-019-1138-y

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.806283

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, Purpose and Teleology. *Philosophy of Science*, *10*(1), 18–24.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. http://aima.cs.berkeley.edu/

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Sen, A. K. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, *6*(4), 317–344.

Serenko, A., & Turel, O. (2022). Directing Technology Addiction Research in Information Systems: Part II. Understanding Technology Addiction. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *53*(3), 71–90. https://doi.org/10.1145/3551783.3551789

Seyfarth, R. M., & Cheney, D. L. (2013). The Evolution of Concepts About Agents. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199890712.003.0006

Simon, H. A. (1979). Rational Decision Making in Business Organizations. *The American Economic Review*, *69*(4), 493–513.

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, *71*(3), 549–570. https://doi.org/10.1037//0022-3514.71.3.549

Spelke, E. S. (2022). Agents. In E. S. Spelke (Ed.), *What Babies Know: Core Knowledge and Composition Volume 1* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780190618247.003.0007

Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Swanepoel, D. (2021). Does Artificial Intelligence Have Agency? In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts* (pp. 83–104). Springer International Publishing. https://doi.org/10.1007/978-3-030-72644-7_4

Tokunaga, R. S. (2017). A meta-analysis of the relationships between psychosocial problems and Internet habits: Synthesizing Internet addiction, problematic Internet use, and deficient self-regulation research. *Communication Monographs*, *84*(4), 423–446. https://doi.org/10.1080/03637751.2017.1332419

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Winner, L. (1977). *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. MIT Press.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 3770. https://doi.org/10.1038/s41467-019-11786-6