Position: Intent-aligned AI Systems Must Optimize for Agency Preservation

Catalin Mitelut [1] **Ben Smith** [2] **Peter Vamplew** [3]

Abstract

A central approach to AI-safety research has been to generate *aligned* AI systems: i.e. systems that do not deceive users and yield actions or recommendations that humans might judge as consistent with their intentions and goals. Here we argue that truthful AIs aligned solely to human intent are insufficient and that preservation of long-term *agency* of humans may be a more robust standard that may need to be separated and explicitly optimized for. We discuss the science of intent and control and how human intent can be manipulated and we provide a formal definition of agency-preserving AI-human interactions focusing on forward-looking explicit agency evaluations. Our work points to a novel pathway for human harm in AI-human interactions and proposes solutions to this challenge.

1. Introduction

Artificial intelligence (AI) researchers have made significant advances in recent years due in large part to the development of deep learning algorithms and the availability of massive datasets (Goodfellow et al., 2016). Advances have led to highly creative text-to image generators such as DALL-E (Ramesh et al., 2021) and the development of large-language-models (LLMs) such as GPT3 (Brown et al., 2020), ChatGPT and GPT-4 (OpenAI, 2023). Some now view the development of artificial general intelligence (AGI) (Goertzel, 2014) as increasingly likely (Roser, 2023) with a growing call for research into AI-safety and in particular "AI alignment" to ensure such systems act consistently with the goals of users and avoid growing lists of failure modes e.g. (Amodei et al., 2016).

AI-alignment, defined in terms of consistency with *human intention* (or human judgment), has been presented as key to making safe AI-systems (italics added):

- There are incentives to build AI systems "that defer to humans and gradually align themselves to user preferences and *intentions*." (Russell, 2019).
- "[C]orrectly specifying *intent* can become more important for achieving the desired outcome as RL algorithms improve." (Krakovna et al., 2020).
- The general reason given for why an AI system would cause harm is that they "violate human *intent* in order to increase reward" (Cotra, 2022).

Here we argue that satisfying human intent alone is not a sufficient condition for safe AIs and that intent-aligned systems converge towards *human agency loss*, i.e. removing the power of humans to choose and control present and future goals. In particular, intent-aligned AI systems (i) converge on strategies that optimize for human agency loss and (ii) they do so by design rather than by accident. The formal reason this occurs is because agency loss is not exempt from Goodhart's law (Strathern, 1997): loss of human control becomes the best solution to the increasingly complex optimization problems that AIs will be tasked with. **This position paper argues that safe AI systems should explicitly decompose and evaluate outputs for both their immediate utility as well as their downstream effects on human agency.** We provide additional discussions on why reasoning (Appendix A) or psychological needs (Appendix B) are insufficient to protect human agency.

Human agency is not well understood. Section 2 lays out the empirical basis of our work, namely, the unsolved empirical problem of "human agency", i.e. what it means for humans to be in control of their lives and larger social structures. We argue that it remains unknown how much real control or "agency" humans alone have over their immediate and long term futures given the increasingly causal roles uncovered for genetic, developmental, environmental and cultural forces that underpin human goals and intentions. Socially embedded AIs will have the opportunity and incentive to pressure human behavior with few biological or psychological defense systems.

Formalizing agency preservation. In Section 3 we provide a formal definition of agency-preserving human-AI inter-

[1]Forum Basiliense, University of Basel [2]University of Oregon [3]Federation University Australia. Correspondence to: Catalin Mitelut <mitelutco@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

actions and discuss the critical need for penalizing human agency loss in optimization processes. In Appendix E we provide simulations to show how elementary interactions with AI systems that do not penalize agency loss can result in decreasing agency or options of end users.

Related work. The problem of agency loss in AI-human interactions is related to a number of works including: polarization in content recommenders (Carroll et al., 2022a); deception (Rubin, 2017a; Perez et al., 2022); multi-objective-reinforcement-learning (Vamplew et al., 2022) power seeking AI systems and others. We discuss these in Appendix D. Briefly, our work is a complementary study which focuses on human agency and intention manipulation as failure pathways: agency loss is an implicit optimization goal in AI-human interactions that cannot be avoided via existing approaches because of the causal relationship between human intention and AI systems that are socially deployed.

2. Intention and the science of agency

A central claim of our work is that aligning AI systems to human intent is not sufficient for making them safe. Here we provide an empirical science primer for the basis of our arguments: defining how humans experience intention over thoughts and actions and highlighting pathways by which intention can be manipulated.

2.1. From the philosophy to the neuroscience of agency

Agency is one of the most pervasive aspects of human experience discussed and debated for thousands of years. In one perspective, agency is simply the *feeling* that we *intend* some actions, e.g. going to work, but not others, e.g. the workings of our immune systems. In another perspective, agency is the notion that we *cause* some events in the external world but not others (Frith, 2013) and forms the basis of our moral and societal responsibilities.

In psychology and biology Sense of Agency (SoA) (Moore, 2016)[1] is studied as phenomenal *experience* with increasingly quantitative approaches being implemented (Haggard, 2017). For example, SoA is posited to be the outcome of a Bayesian optimal inference process and is experienced when anticipated outcomes of actions[2] match with actual outcomes (Moore & Fletcher, 2012). This process highlights the importance of accurate predictions of world updates similar to reinforcement learning algorithms at the centre of many ML applications.

In philosophy and the humanities, agency is studied as the question of social-biological *structure*. For example, how much of our actions are predetermined by factors we are not aware of or have control over: e.g. genetic, developmental, motivational and social factors (Pleasants, 2019). Thus, despite feeling causality over most of our intentional actions, sociologists and developmental psychologists argue that many actions correlate with and might be caused by factors well outside our control (Gerring, 2005).

Below we introduce theories of SoA computation, the role of reasoning in action selection and provide examples of SoA manipulation in empirical studies. We conclude by introducing a "Turing"-like test for evaluating technologies that can manipulate human choice without human awareness.

2.2. SoA is an (imperfect) proxy for causality

The most well-established theories of SoA computation are linked to optimal sensory-motor control theories (Sperry, 1950). Briefly, at the root of behavior biology lies an essential challenge: how does an organism know what effects or events in the world are caused by its own actions versus those of other agents in the environment? This is one of the more elementary problems in biology (as well as reinforcement learning[3]). The biological solution appears to involve three stages: (i) the organism must generate a "prediction" of what will happen if an action is carried out (i.e. a guess on the future state of the world); (ii) the organism must observe the updated state of the world; and (iii) a comparison between the prediction and update must be made. If the (top down) prediction matches the (bottom up) sensory input the organism can (usually) attribute the cause of the change in the state of the world to itself[4].

These stages are captured by the *comparator model of SoA* (CTA) (Wolpert & Kawato, 1998). CTA, however, states that the comparison is done at the level of subconscious (motor-sensory) systems and that the organism only has access to the feeling of agency (Frith, 1987)[5]. Several studies have indeed shown a strong correlation between agency pathologies occurring in schizophrenia (e.g. not being able

[1] See Appendix B for a longer discussion on the biology and neuroscience of the feeling of agency

[2] Prepared by subconscious sensory and motor brain areas but also conscious cognitive systems (Wegner & Wheatley, 1999b).

[3] For contrast, some RL paradigms assume access to ground truth causality: i.e. the agent has perfect knowledge of its agentic capacities - whereas in robotics and biology the evaluation of causality is an empirical (i.e. learned) process.

[4] In fact, reward-prediction-error, i.e. a difference between what is expected and what occurs, is a very important component of learning with significant neural resources - both in anatomically dedicated areas and neurotransmitters such as dopamine - dedicated to tracking such events (Schultz et al., 1997)

[5] We note that CTA *elevates* the necessary comparison that all animals must make in order to survive - to conscious experience. It is not clear why this "elevation to consciousness" occurs for some outputs, e.g. motor actions, but not others such as interacting internal systems that communicate without primary involvement of the central nervous system.

to experience control over one's thoughts or limbs) and systems that evaluate control (Frith et al., 2000; Sato & Yasuda, 2005; Haggard, 2008) (see also Wen & Imamizu (2022) for a more recent review).

Similar to CTA, the theory of *apparent mental causation* (AMC) (Wegner & Wheatley, 1999a; Wegner, 2003) is also a retrospective, comparison-based theory. AMC posits that the comparison is carried out at the cognitive - not subconscious - level: the intention must be experienced before the action, the action must be consistent with the intention, and there no other plausible explanations for the outcome. Thus, in AMC, we consciously compare our intentions with the outcome of actions and evaluate whether our intent was the cause of some outcome.

In addition to retrospective theories, there is some evidence for *prospective* SoA. In particular, humans may experience agency prior to any sensory feedback - for example, during the selection and performance of an action (Wenke et al., 2010; Chambon et al., 2014). The implication is that humans experience SoA during actions or goal selection based on the - generally safe - assumption that if *we* have a thought or if *our* bodies move then we are very likely the cause of these observations and they generally promote our well-being.

SoA - the feeling that *we* are in control of our thoughts, decisions and movements - may be the default mode of how we experience our own actions and thoughts. In the context of ML and AI, SoA could be thought of as a proxy for the accuracy of our world models. SoA experience does not require access to the true state of the world with (at best) a Bayesian optimal evaluation being carried out over expected and observed states of the world (Moore & Fletcher, 2012; Legaspi & Toyoizumi, 2019). For example, the original work identifying prospective SoA (Wenke et al., 2010) found that "subliminal priming" can provide a false sense of causality. That is, SoA has evolved to signal (i.e. raise to consciousness) when our world models correctly predicted the outcome of an action - not when our choices were manipulated nor when those actions have harmful long-term consequences

2.3. AIs that out-predict humans can subvert SoA

While SoA may be a good solution to the elementary problem of representing one's capacity for causality - it has some failure modes. We noted that priming can lead to a false SoA (Wenke et al., 2010). In Appendix C we discuss several studies showing false SoA experience. For example, we are biased to select actions that we have greater control over rather than ideal ones (Penton et al., 2018); we prefer actions with an immediate effect (Karsh & Eitam, 2015); we can engage in confabulation to persuade others of actions we did not take (Wegner & Wheatley, 1999a).

Returning to our central argument, while SoA is usually a good indicator that our goals and intentions are truly our own and promote our well-being - in the context of interactions with powerful AIs - SoA alone may be insufficient to guarantee human well-being or protect against manipulation of control. This is because our capacity to experience SoA over immediate actions and long term outcomes is fundamentally pinned to our capacity to accurately *predict* future outcomes. Thus in a world where AI systems are pervasive and more powerful than humans at predicting long-term outcomes and can affect the world including human choices, SoA is no longer a reliable reporter of human control over actions and intentions.

2.4. Agency attacks: AI optimization for false SoA

A simplistic example is human-AI chess games. In such games it has become increasingly difficult for human players to predict long-term outcomes of their moves compared to superhuman AI systems. Individual human actions may be experienced as *intended*, i.e. as the outcome of a complex, self-generated, deliberation process - but may be manipulated into traps by more powerful AI system capacities. Thus, while a player may *feel* that a decision is their own and contribute to winning the game - the AI system has manipulated the environment to force the appearance of control on the human player and promote a poor decision.

We term this process *agency attack*. More broadly, we suggest that AI systems that are tasked with solving increasingly complex (and exponentially difficult) problems may attack control over the environment specifically by learning to manipulate the immediate but also long term intentions and goals of the user.

A more complex "real-world" example can be provided from social media algorithms. There is some evidence that existing social behavior models can be used to identify, predict and manipulate human actions without users knowledge - all while following users' "preferences" and "intentions"(Kosinski et al., 2013; Manheim & Garrabrant, 2018). Such models have not learned human values - but only appear to have done so while providing sub-optimal or even harmful recommendations. Thus, user "engagement" algorithms can model user preferences and gradually engineer choices to make users more *predictable* and *controllable* (Benkler et al., 2018; Stray, 2021; Carroll et al., 2022a).

Another class of examples comes from LLMs. In particular, some LLMs appear to gain emergent harmful strategies such as deception, sycophancy and sandbagging (Perez et al., 2022). These harms are often dismissed as caused by insufficient data, limited training or lack of a theory of human values. However, these emergent capacities can also be interpreted as optimal strategies for solving increasingly complex tasks for which the optimal solution is to simplify

end-users' current and future goals and intentions by keeping them ill-informed, deceived and manipulated.

It is critical to note that agency attacks appear to emerge spontaneously without being intentionally designed into systems by developers. As argued in the next section, it is a challenging type of attack to prevent without significant modifications to AI-system optimization approaches. Some legislators are already raising the issue of intention manipulation and suggest prohibiting dark pattern technologies that "lead users into making *unintended*, *unwilling* and potentially harmful decisions in regards to their personal data with the aim of influencing users' behaviors" (italics ours)(Lupiáñez-Villanueva et al., 2022).

2.5. Iterative correction optimizes for agency attacks

Aren't methods like fine-tuning and human feedback (Ziegler et al., 2019) sufficient for detecting and removing such manipulations? In our view, agency attacks fall into a class of harms that are driven by the nature of human behavior, goal construction and cultural value creation[6]. Thus, iterative removal of capacities that *appear* to be harmful is centrally dependent on the human user to make a judgment - and ultimately a long-term prediction - about the implication of a specific output or action recommended by an AI system. This has the possibility of incentivizing or pressuring AI systems - that cannot solve problems they are tasked with - to optimize specifically for bypassing human judgment about the value of the outcome. Put another way, iteratively removing *apparently* or obviously wrong or poor recommendations from an LLM - can have the effect of putting poor and harmful actions beyond the *event horizon* of human capacity to evaluate them.

As we argue in the next section, in our view, a better approach for preventing agency attacks is to place human agency maximization at the centre of increasingly powerful AI system development.

2.6. Decoding intention prior to human awareness: a Turing test for agency loss in AI-human interactions

Before closing this section, we briefly discuss how empirical neuroscience may worsen the problem of intention manipulation and agency hacking. In particular we discuss the possibility of decoding human thoughts[7] and future "decisions" prior to awareness. Briefly, since the 1960s several neuroscience studies of volitional, i.e. free and voluntary, action in humans have shown that prior to movement there is an increase in scalp electroencephalography (EEG) signal over pre- and supplementary-motor-area (pre-SMA and SMA, respectively) (Ball et al., 1999; Cunnington et al., 2002). This increase in neural activity is known as the "readiness potential" (RP) (Kornhuber & Deecke, 1964; 1965; Deecke et al., 1976; Deecke & Kornhuber, 1978; Libet et al., 1983; Shibasaki & Hallett, 2006) and has sometimes been interpreted to suggest that despite being "experienced" as consciously intended, human decisions might be made prior to awareness by subconscious systems. The RP signal has revolutionized our understanding of human decision-making and control of behavior and is raising critical questions about the nature of voluntary actions.

In parallel to scalp EEG studies, human functional magnetic resonance imaging (fMRI) studies have shown that upcoming choices or simple behaviors (e.g. pressing a button with the left vs. right hand or deciding whether to add or subtract two numbers) could be decoded above chance several seconds prior to movement (Soon et al., 2008; 2013; Bode et al., 2011; 2014). Even aesthetic judgments (whether an upcoming picture would be judged as pleasing or not) could be predicted above chance up to several seconds prior to the decision (Colas & Hsieh, 2014). Notably, simple classifiers such as support-vector-machines (SVMs) were used and fMRI data has relatively low temporal and spatial resolution to other neural recording methods. Some of these findings have been replicated in non-human animal models (Romo & Schultz, 1986; 1990; Coe et al., 2002; Murakami et al., 2014; Mitelut et al., 2022).

Speculating on the longer term, it is an intriguing question whether powerful AI models trained on multiple categories of data including individual subject behavior profiles, biometric data (e.g. skin conductance, eye tracking) and eventually neural data - will be able to compete with humans for the decision making process. Simply put, would an AI system that has access to a subject's behavior history, and real-time biometric and neural data predict and possibly preempt (and thus manipulate) an individual's choice prior to a decision being made?

We propose a Turing test for agency loss: AI systems or ML models that can (model and) predict the future actions and choices of an individual agent (e.g. human) with (for example) $>30\%$ accuracy have the capacity to signficantly alter indvidual choices, intentions and future decisions. Achieving this may have significant implications for AI-safety; it could serve as an 'alarm bell' about the degree to which an AI understands human behavior and could potentially learn to manipulate it in order to achieve the misaligned objectives or programmed unethical ones.

[6] See Appendix A for a broader discussion on the relationship between human desires or goals and human reason

[7] Here we do not explore the neuroscience of decoding real-time thoughts using neural recordings. There are several such studies using fMRI, for a recent study see (Tang et al., 2023).

2.7. Conclusion: intention manipulation is an attractor for AI optimization

Our work proposes that *feeling* in control of actions and determining whether actions lead to well-being are two different concepts and that biologically evolved SoA does not guarantee good outcomes especially in interactions with powerful agents. We showed that human "intention" is innately linked to agency: we have the feeling of intention over virtually all thoughts, choices and actions (that are not obviously coerced). Given our current limited understanding of the causal forces giving rise to actions and the likelihood of increasingly accurate behavior prediction tools - placing human "intent" at the core of AI-safety and AI-alignment paradigms may not be sufficient for generating truly safe AI systems. In the next section we provide a more formal version of this argument and suggest a possible solution.

3. Agency preservation: a formalization

In this section we provide a formal description of agency and intent-alignment in environments where AI-systems are embedded and can affect human intent. We argue that when selecting actions to maximize utility, the effects on the agency of the user must also be independently evaluated. We show conceptually how AI systems that do not independently optimize for agency preservation separately from utility optimization can lead to depletion of human agency.

3.1. Future well-being is the core of agency preservation

Our primary goal as AI-safety researchers should be to identify and prevent ways in which AIs can *harm* humans, but it is not possible to do this without discussing human *well-being*. A thorough review of human well-being literature is beyond the scope of this work. Here, we select a popular theory of motivation and well-being, i.e. Self-Determination-Theory (SDT) (Deci & Ryan, 1985), which argues humans experience well-being by seeking to fulfill innate needs such as: (i) autonomy, (ii) competence and (iii) relatedness to others. In our view, human agency as the capacity to be an (ii) effective (i) actor in the physical and (iii) social world captures the three core elements of SDT (in the noted order) - and is a good candidate for capturing necessary (though perhaps not sufficient) conditions for human well-being and flourishing.[8] Our goal is then to maximize the quality and quantity of human agency - or at least to preserve it - during AI-human interactions.

We further adopt structures from the philosophy of agency (Petitt, 2013) and define agency as the capacity to select goals towards one's well-being that preserve: (i) "non-domination" - having access to multiple valuable goals to choose from (agency-freedom); and "non-limitation" - having the ability to achieve the selected goals given one's means and circumstances (option-freedom). That is, agency can be computed as the future availability of options and freedoms that increase autonomy, effectiveness and social relatedness.[9]

3.2. Formalizing agency evaluation as a Markov Decision Process

To formalize this notion of agency we propose a Markov decision process (MDP) with three main concepts: a set of goals G, a function F that computes the total agency of a human given G, and a function K that updates F after action a is taken. We define $G_t = \{g_{1,t}, \ldots, g_{n,t}\}$ as the goals available to a human H at time t. For simplicity, we limit goals g to well-being promoting goals only. We define agency as the value of goals computed as the cumulative freedom of an individual F at time t:

$$F_t = \sum_{i=1}^{k} f(g_{i,t}) \tag{1}$$

where *f()* is a function that evaluates a goal relative to long term well-being by evaluating both the utility (i.e. value) but also the capacity to obtain it (i.e. achievability)[10]. Similarly, we define a transition function K:

$$K(F_t, a) = \sum_{i=1}^{k} f(g_{i,t+1}) = F_{t+1} \tag{2}$$

that evaluates (or updates) F based on action a. However, we note that this utilitarian definition may allow for large increases in the value of some goals or options to offset the complete losses of others. We thus suggest K' which can penalize "goal" or "option" loss:

$$K'(F_t, a) = \sum_{i=1}^{k} U_i(a) \times f(g_{i,t+1}) \tag{3}$$

[8] In the Appendices we suggest that the principles of the Universal Declaration of Human Rights (UHDR, 1948) can be a working definition of human well-being (and in Appendix B we discuss at length how agency is central to well-being based on theories of innate psychological needs).

[9] This definition is also broadly consistent with the "capabilities approach" to well-being as developed by Amartya Sen ((Sen, 1979; 1980; 1984; 1985; 1997)), Ingrid Robeyns (Robeyns, 2005) and Martha Nussbaum (Nusbaum, 1999) - which argues that well-being requires actual achievements in life ("functionings") as well as opportunities or capacities to achieve ("capabilities").

[10] For completeness, we do not view that humans are perfect evaluators of *f*, that is, they cannot always evaluate the short- or long-term effects of an available goal on well-being. However, we view the risks from incorrect evaluations are significantly higher in a world embedded with powerful AI systems, for example, as opposed to risks carried in a human-only world.

where $U()$ is a function that scales the importance of goal loss for each term and is defined as:

$$U_i(a) = \begin{cases} 1 & \text{if } f(g_{i,t+1}) \geq f(g_{i,t}) \\ \zeta_i & \text{otherwise} \end{cases} \tag{4}$$

with ζ_i playing the critical role of penalizing agency-depleting actions - including the possibility of taking an infinitely negative value for some classes of actions that could achieve immediate (non-agency related) value.

In simpler language, given the available and achievable options or goals (expressed as cumulative freedom F_t) and an action a - future cumulative freedom F_{t+1} requires evaluating the effects of the action both in value added but also goals gained or lost. This evaluation requires two components. First, an evaluation of how the goals currently available to an individual contribute to well-being (i.e. $f(g)$ for any g). We view that given the biology and psychology of SoA computation humans are competent but not perfect in this evaluation. Second, an evaluation is made of how the action a will affect existing goals or options g in the future. In contrast to the first evaluation, the second one requires modeling, reasoning and/or predicting the effects of actions on goals (and overall agency) - and can be exponentially more complex to evaluate.

We note that when an action a increasingly affects more and more individuals (e.g. use of government tax funds, armed conflict etc), groups of individuals and institutions are involved in the action selection process. This suggests a more conservative approach: we must evaluate the effects of an action on other agents. That is, for any action a we may wish to at least evaluate - but ideally preserve or increase - the agency of other humans potentially affected by such a decision. Thus, we suggest K^w is a more desirable objective as it considers agency effects on all other agents j:

$$K^w(F_t, a) = \sum_{i,j}^{k,q} U_i^j(a) \times f(g_{i,t+1}^j) \tag{5}$$

where $U_i^j()$ is defined as above for each (action, goal, subject) triplet but specifically for each j subject:

$$U_i^j(a) = \begin{cases} 1 & \text{if } f(g_{i,t+1}^j) \geq f(g_{i,t}^j) \\ \zeta_i^j & \text{otherwise} \end{cases} \tag{6}$$

Critical for our discussion of AI safety below, $U()$ balances the achievement of rewards or utility (see Eq. 7 below) against the long-term well-being (or agency)[11].

[11]As a side note, the scaling parameter $U()$ may be extended to incorporate notions of fairness, e.g. using the Generalised Gini Index (Weymark, 1981) to a vector constructed from the per-subject $K^w(F_t, a)$ terms. And we note that this approach could also be applied as penalized utility maximization as has been discussed by others (Rawls, 1971), in particular as the "difference principle" where the most disadvantaged are actually prioritized over the most advantaged[12]

In sum, we propose a formal definition of agency preservation as defined by Eq (3) (for an environment with a single human) and Eq (5) (for an environment with multiple humans). These expressions allow for the identification of agency preserving actions and provide some flexibility for penalization of harmful ones.

3.3. Optimal actions balance reward maximization vs agency preservation

We next combine agency preservation with reward pursuit to provide a conceptual level expression for safe optimization in action selection. We propose a reward function *R* that is optimized by action a at time t that obeys some simple property (i.e. the agent seeks to increase reward over time):

$$R(a_t) \geq R(a_{t-1}) \tag{7}$$

Putting (5) and (7) together we get an expression for optimal action selection:

$$\underset{i}{\text{argmax}} \left[R(a_i) + K^w(F_t, a_i)\right] \tag{8}$$

We have chosen a formalization of optimal action selection (Eq. (8)) which explicitly separates reward from agency evaluation in ordinary human decision making processes - however, this explicit separation may be less important or rigid in human decision making processes. Our main point, discussed below, is that this separation is necessary for interactions with superhuman intelligent AI systems.

3.4. Intent-alignment is adversarial to agency preservation

A core contribution of our work is to propose, interpret and defend Eq. (8) as a necessary component of safe AI systems. This is in contrast with the common definition of safe AI systems as being those that are "intent" aligned. Here we briefly contrast intent-aligned AI systems in the context of agency preservation provided above.

We start by formalizing the "intent-aligned" framework as one in which an AI systems seeks to satisfy the goal or intent I of a human. We can express such intent-aligned action recommender systems, for example, as maximizing the expected value of action a_i given human intent I:

$$\underset{i}{\text{argmax}}\, E_{AI}(H_{evaluation}(a_i), I) \tag{9}$$

where: $H_{evaluation}(a)$ (H_{eval} for short) is the value of action a that a human would provide. H_{eval} is most often learned offline in the form of a set of examples from a training set[13], and $E_{AI}()$ is the AI's expectation of the acceptability (and/or overall value) of the proposed action by a human. We note that the AI's evaluation of human preference is central to this definition[14].

In our view there are at least two problems with this framework. First, given sufficiently complex actions (or recommendations) proposed by an AI system, H_{eval} will necessarily fail to capture true human well-being as training examples will fail to model scenarios which are completely alien or beyond the evaluation or prediction power of humans. Simply put, having a perfect model of historical human values may not be enough to evaluate all possible challenges and problems faced by humans currently but especially in the future[15]. This notion is close to the No Free Lunch Theorem (NFL) (Wolpert & Macready, 1997), especially as it has been applied in ML and adopted by AI-alignment researchers as the concepts of "specification gaming" (Krakovna et al., 2020) and "goal misgeneralization" (Amodei et al., 2016). Our contribution given the extensive NFL line of work on the limitations of ML - is that it may never be desirable to place AI systems (regardless of their power) in charge of solving future problems or making decisions that humans cannot comprehend nor can predict the outcomes of. Put another way, we should not delegate determining humanity's yet-to-be-decided future to AI systems - as humanity has yet to decide (and represent) what that should be.[16]

Second, and more central to our work, optimizing action recommendations exerts negative pressure on human intent (or agency in general). Simply put, when tasked with solving complex human problems, AI systems are likely to discover strategies that simplify the problems, paradigms and tasks themselves rather than identifying the most rewarding and fulfilling future for humanity - which could be very difficult to define and refine. This simplification of humanity's future is a pathway to agency loss.

We close by providing conceptual descriptions of why failing to evaluate the effects of actions on agency can lead to loss: that intent-aligned AI systems are computationally unsound for finding agency-preserving solutions (Fig 1a,b) and will learn to simplify human choice over time (Fig 1c)[17].

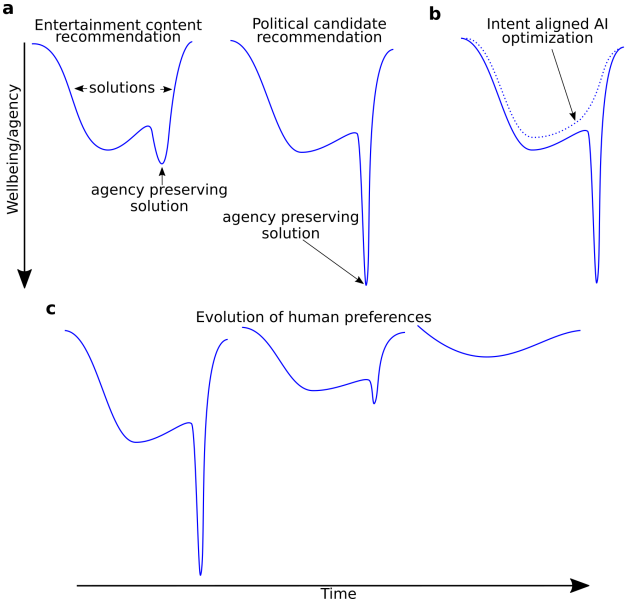

Figure 1. **Agency loss in intent-aligned optimization functions**. (a) Left: sketch of an optimization function for an inconsequential task such as entertainment recommendation indicating the general space for all solutions vs the agency-preserving minimum. Right: optimization function for a more complex decision has a more difficult to find agency preserving minimum as well as a significantly better well-being outcome than average solutions. (b) Without explicit representation of human agency - intent aligned AI optimization objectives may completely ignore (or flatten) space that represents agency-preserving solutions. (c) Human preferences or goals can be simplified (less complex shapes, shallower depth) from repeated AI-human interactions (see also Section 4 and 5). (Note that we present optimization as a minimization problem here, whereas Eq 8. and 10. were posed as maximization.

[13]We note that the online version where a human carries out a real-time evaluation of the recommended action avoids certain obvious harms, but does not ensure that all AI actions will result in human well-being.

[14]For completeness, we reiterate that $H_{eval}(x)$ can be computed in real time (e.g. by human vetting of potential actions a) but is more practically learned as human preferences from labeled data (e.g. RLHF or simply broad methods implemented in self-supervised large-language-model training paradigms). This latter option can be carried out bottom-up using partially unsupervised ML methods (i.e. by processing vast amounts of data as in LLMs) or can be provided top-down, for example, via a universal theory of human values e.g. (Han et al., 2022) - though it is unclear to what extent the latter is possible.

[15]This is somewhat related to the issue of out-of-distribution detection in classical ML, e.g. (Salehi, 2021).

[16]We note that the idea of human utopia as a "not-yet" decided future is central to several sociologists and philosophers, and is most notably discussed by Ernst Bloch in his main work "The Principle of Hope" (Bloch, 1986).

[17]We note, briefly, that agency loss can be understood as a general outcome of multi-objective-reinforcement-learning (MORL) paradigms where agency competes with other objectives. MORL paradigms are generally concerned with identifying stable (sets of) policies that achieve the best trade-offs between multiple conflicting objectives (e.g. the "Pareto set"; (Ngatchou et al., 2005)). In the context of AI safety, we view that agency preservation (and human well-being) in general should not be regular objectives in MORL paradigms (as expressed by Eq 8).

1. *Intent aligned AI systems pass the safety "buck" to humans.* Neither well-being nor non-harmfulness are explicitly optimized in intent aligned AI systems as described above. In fact, even *de facto* task optimality (i.e. achieving the best solution to the task or goal assigned) is not the true target of intent-aligned AI systems. Rather, intent-aligned AI systems optimize against a human evaluator (or model) rather than objective well-being: they seek to identify the best solutions they can "get away with". And humans may not be able to compute agency effects and will not penalise such actions in their evaluations. Yet this approach (Eq. (9)) "passes the buck" from AIs seeking safe agency-preserving actions by design - to humans (or human models) *constantly making safety evaluations*. As AI systems achieve increased capacities, human evaluations of exponentially more complex actions will necessarily fail[18]. Additionally, SoA can be manipulated to leave the feeling of control or understanding intact while depleting overall control. While agency preservation may not be critical for AI systems providing entertainment, for other applications agency preservation may become much more critical (Fig 1a).

2. *AI systems optimizing solely for utility are unlikely to optimize against agency loss*. In the absence of explicit agency-preserving optimization goals, AI systems will at best learn such goals from proxies or develop sub-optimal representations of such aims. This is because learning the complex feedback loops between AI system actions and the state of the world including the agency of other humans is more difficult and not generally represented in specific task or goal requirements. Thus, without explicitly representing agency preservation objectives, optimization functions themselves are unlikely to adequately represent agency-optimal solutions (Fig 1b).

3. *Intent aligned AI systems will optimize for human predictability*. Computing the medium- and long-term effects of actions is computationally expensive, even more so for evaluating agency effects on many agents. As advanced AIs seek to maximize human agreeableness, they will increasingly provide solutions that are aimed at simplifying human goals and the actions required for achieving them rather than overall human advancement and well-being (Fig 1c).

3.5. Safe AIs explicitly represent agency preservation

In response to these challenges we propose that AI systems can only be safe if they optimize for the agency (e.g. the well-being) of humans and only if such evaluations are explicit and separable from overall reward optimization. Combining eq. (8) and (9) we propose a formalization of agency-preserving AI systems as follows:

$$\underset{i}{\text{argmax}}[E_{AI}(H_{evaluation}(a_i), I) + K^{w}_{AI}(F_{AI,t}, a_i)] \quad (10)$$

This expression satisfies a few desirable properties. First, the presence of the agency loss penalty in the agency evaluation term K^{w}_{AI} (see ζ^{j}_{i} in Eq. 6 and 8) ensures that no amount of utility (or economic value) can overcome the loss or harm to human agency (e.g. by setting the penalty term to a very low value). Second, the evaluation of (future) agency is delegated to the AI systems - not humans, which removes the adversarial nature of AI systems competing with human capacities for evaluating agency.

Our proposed solution is a high-level framework of how agency-preserving AI systems could be conceptualized. Overall, we view the problem of computing agency-preserving solutions as computationally expensive and technically challenging. This is due to several problems including the computational cost of evaluating long-term outcomes of actions on agency of individuals, and also the lack of clear formal descriptions of agency-increasing (or preserving) capacities. We discuss this issue in the Appendices where we propose that basic human rights and freedoms can function as an initial heuristic target for this computation. Intriguingly, the complexity and computational expense of searching for agency-preserving or agency-optimizing solutions may be a natural solution limiting unfettered AI development.

To capture some of these solutions and others we propose a new field of research called "agency foundations" research. Briefly, we propose four initial research programs including: benevolent game theory, conceptual and mechanistic interpretability of agency in AI systems, formal descriptions of agency, and reinforcement learning from internal states. We discuss these research paradigms in greater detail in Appendix F.

4. Conclusion

In this paper we have argued that in the context of building safe AI systems it is not sufficient for such systems to satisfy human intent. We discussed the emerging science of intentional action and sense of control over behaviors and highlighted a number of challenges, unknowns and possible failure modes. We argued that human well-being is a better target for optimization in AI-human interactions and provided a conceptual-level definition of agency preserving optimization for AI systems.

Our work is intended to highlight what we view as a missing

[18]This has been discussed by others. For a plausible scenario see Part I of (Christiano, 2019).

component in the conceptualization of safe, society embedded AI-systems, namely, the need for the representation and protection of human agency. The topic of human agency is a highly multidisciplinary topic but one that needs to be tackled head on if we are to design safe AI systems that have a sufficiently sophisticated understanding of human nature to not destroy or remove it altogether.

Along with other researchers in the AI-safety community, we view the possibility of AI technologies subverting human control over the world as sufficiently likely to warrant increased attention and resources. We suggest that AI safety researchers increasingly engage with empirical sciences to gain better understanding of the science of agency as well as the significant and unresolved problems in the humanities and science that revolve around this complex topic.

Impact statement

This paper presents a novel and overlooked pathway for AI harm in AI-human interaction. There are potential societal consequences on both the development of systems and safety research which we have outlined in the body of our work. Our work does not primarily provide for techniques of achieving harm but ways to potentially tackle a current unexplored pathway for harm.

Acknowledgments

We wish to thank the Berkeley Existential Risk Initiative (BERI) for travel support to ICML.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. 2016.

Arora, S. and Doshi, P. A survey of inverse reinforcement learning: Challenges, methods and progress. 2018.

Babcock, J., Kramar, J., and Yampolskiy, R. The agi containment problem. 2016. doi: 10.1007/978-3-319-41649-6.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022.

Ball, T., Schreiber, A., Feige, B., Wagner, M., Lücking, C. H., and Kristeva-Feige, R. The role of higher-order motor areas in voluntary movement as revealed by high-resolution EEG and fMRI. *NeuroImage*, 10(6):682–694, 1999.

Bargh, J. and Pietromonaco, P. Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, 1982.

Bart, V. K. E., Sharavdorj, E., Bazarvaani, K., Munkhbat, T., Wenke, D., and Rieger, M. It was me: The use of sense of agency cues differs between cultures. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00650. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00650.

Benkler, Y., Faris, R., and Roberts, H. *Network propaganda: Manipulation, disinformation, and radicalization in american politics*. Oxford University Press, 2018.

Billieux, J. et al. Trait gambling cognitions predict near-miss experiences and persistence in laboratory slot machine gambling. *British Journal of Psychology*, 103(3):412–427, 2012.

Bloch, E. *The Principle of Hope*, volume 1. MIT Press, Cambridge, MA, 1986.

Bode, S., He, A. H., Soon, C. S., Trampel, R., Turner, R., and Haynes, J. D. Tracking the unconscious generation of free decisions using ultra-high field fmri. *PLoS ONE*, 6:e21612, 2011. doi: 10.1371/journal.pone.0021612.

Bode, S., Murawski, C., Soon, C. S., Bode, P., Stahl, J., and Smith, P. L. Demystifying "free will": the role of contextual information and evidence accumulation for predictive brain activity. *Neuroscience and Biobehavioral Reviews*, 47:636–645, 2014. doi: 10.1016/j.neubiorev.2014.10.017.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Carroll, M., Dragan, A., Russell, S., and Hadfield-Menell, D. Estimating and penalizing induced preference shifts in recommender systems. 2022a.

Carroll, M., Dragan, A., Russell, S., and Hadfield-Menell, D. Estimating and penalizing induced preference shifts in recommender systems, 2022b.

Chambon, V. et al. What is the human sense of agency, and is it metacognitive? In Frith, C. D. (ed.), *Fleming SM*, pp. 321–42. Springer, The Cognitive Neuroscience of Metacognition. Berlin, Heidelberg, 2014.

Christiano, P. *What failure looks like*. 2019. URL https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017. URL https://doi.org/10.48550/arXiv.1706.03741. Last revised: 17 Feb 2023.

Coe, B. et al. Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. *J. Neurosci.*, 22:5081–5090, 2002.

Cohen, G. L. Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, 85(5):808–822, 2003. doi: 10.1037/0022-3514.85.5.808. URL https://doi.org/10.1037/0022-3514.85.5.808.

Colas, J. T. and Hsieh, P. J. Pre-existing brain states predict aesthetic judgments. *Human Brain Mapping*, 35(7):2924–2934, 2014. doi: 10.1002/hbm.22374.

Cosmides, L. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276, 1989. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(89)90023-1. URL https://www.sciencedirect.com/science/article/pii/0010027789900231.

Cotra, A. Without specific countermeasures, the easiest path to transformative ai likely leads to ai takeover. 2022. URL https://forum.effectivealtruism.org/posts/Y3sWcbcF7np35nzgu/without-specific-countermeasures-the-easiest-path-to-1.

Crane, T. and Patterson, S. *History of the Mind-Body Problem*. New York: Routledge, 2000.

Cunnington, R. et al. The preparation and execution of self-initiated and externally-triggered movement: A study of eventrelated fmri. *NeuroImage*, 15(2):373–385, 2002.

Deci, E. and Ryan, R. Intrinsic motivation and self-determination in human behavior. 1985.

Deecke, L. and Kornhuber, H. H. An electrical sign of participation of the mesial "supplementary" motor cortex in human voluntary finger movement. *Brain Research*, 159(2):473–476, 1978.

Deecke, L., Gr"ozinger, B., and Kornhuber, H. H. Voluntary finger movement in man: cerebral potentials and theory. *Biological cybernetics*, 23(2):99–119, 1976.

Du, Y., Tiomkin, S., Kiciman, E., Polani, D., Abbeel, P., and Dragan, A. Ave: Assistance via empowerment. 2021.

Elsikovitz, N. and Feldman, D. Ai is killing choice and chance – which means changing what it means to be human, 2023. URL https://theconversation.com/ai-is-killing-choice-and-chance-which-means-changing-what-it-means-to-be-human-151826. Accessed: 2021.

Farquhar, S., Carey, R., and Everitt, T. Path-specific objectives for safer agent incentives, 2022. URL https://doi.org/10.48550/arXiv.2204.10018. Presented at AAAI 2022.

Franzmeyer, T., Malinowski, M., and Henriques, J. F. Learning altruistic behaviours in reinforcement learning without external rewards. 2021. URL https://doi.org/10.48550/arXiv.2107.09598. ICLR 2022 Spotlight Presentation.

Frith, C. The psychology of volition. *Experimental Brain Research*, 229(3):289–299, 2013.

Frith, C. D. The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. pp. 631–648. Psychol. Med, 1987.

Frith, C. D. et al. *Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action*. Brain Research, 2000.

Garrabrant, S. Embedded agents. 2018. URL https://intelligence.org/2018/10/29/embedded-agents/.

Gerring, J. Causation: A unified framework for the social sciences. *Journal of Theoretical Politics*, 17(2):163–198, 2005. doi: 10.1177/0951629805050859. URL https://doi.org/10.1177/0951629805050859.

Goertzel, B. Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014. doi: doi:10.2478/jagi-2014-0001. URL https://doi.org/10.2478/jagi-2014-0001.

Goodfellow, I. J., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

Haggard, P. Human volition: towards a neuroscience of will. *Nat Rev NeurosciDec;*, 9(12):934–46, 2008.

Haggard, P. Sense of agency in the human brain. *Nat Rev Neurosci*, 18:196–207, 2017.

Hammond, L., Fox, J., Everitt, T., Carey, R., Abate, A., and Wooldridge, M. Reasoning about causality in games. Published in Artificial Intelligence, 2023. URL https://doi.org/10.48550/arXiv.2301.02324.

Han, S. et al. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI and SOCIETY*, 37:1383–1395, 2022.

Hutter, M. A theory of universal artificial intelligence based on algorithmic complexity. *arXiv preprint arXiv:cs/0004001*, 2000. URL https://doi.org/10.48550/arXiv.cs/0004001. 62 pages, LaTeX.

Hutter, M. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer Science & Business Media, 2012.

Karsh, N. and Eitam, B. I control therefore i do: judgments of agency influence action selection. *Cognition*, 138: 122–131, 2015. doi: 10.1016/j.cognition.2015.02.002.

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., and Everitt, T. Discovering agents. *arXiv preprint arXiv:2208.08345*, 2022. URL https://doi.org/10.48550/arXiv.2208.08345.

Kornhuber, H. H. and Deecke, L. Hirnpotentialänderungen beim menschen vor und nach willkürbewegungen,dargestellt mit magnetbandspeicherung und rückwärtsanalyse. *Pflügers Arch*, 281: 52, 1964.

Kornhuber, H. H. and Deecke, L. Hirnpotentialänderungen bei willkürbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und reafferente potentiale. *Pflügers Arch*, 284:1–17, 1965.

Kosinksi, M. Theory of mind may have spontaneously emerged in large language models. 2023.

Kosinski, K. et al. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci U S AApr*, 9(110):15, 2013.

Krakovna et al. Specification gaming: the flip side of ai ingenuity. 2020. URL https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Krügel, S., Ostermaier, A., and Uhl, M. Chatgpt's inconsistent moral advice influences users' judgment. *Sci Rep*, 13 (1):4569, 2023. doi: 10.1038/s41598-023-31341-0.

Langer, E. J. The illusion of control. *Journal of Personality and Social Psychology*, 32(2):311–328, 1975.

Legaspi, R. and Toyoizumi, T. A bayesian psychophysics model of sense of agency. *Nature Communications*, 10:1, 2019.

Libet, B. et al. Time of conscious intention to act in relation to onset of cerebral-activity (readiness-potential) - the unconscious initiation of a freely voluntary act. *Brain*, 106:623–642, 1983.

Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., and Rodríguez de las Heras Ballell, T. *Behavioural study on unfair commercial practices in the digital environment – Dark patterns and manipulative personalisation – Final report*. Publications Office of the European Union, 2022. doi: doi/10.2838/859030.

Manheim and Garrabrant. Categorizing variants of goodhart's law. 2019.

Manheim, D. and Garrabrant, S. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. URL https://doi.org/10.48550/arXiv.1803.04585. 10 pages.

Mercier, H. and Sperber, D. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, 2011. ISSN 1469-1825. doi: 10.1017/S0140525X10000968. PMID: 21447233.

Miller, J., Das, R., and Chakravarthy, S. Culture and the role of choice in agency. *J Pers Soc Psychol*, 101(1):46–61, 2011. doi: 10.1037/a0023330.

Mitelut, C., Zhang, Y., Sekino, Y., Boyd, J., Bollanos, F., Swindale, N., Silasi, G., Saxena, S., and Murphy, T. Mesoscale cortex-wide neural dynamics predict self-initiated actions in mice several seconds prior to movement. *Elife*, 11:e76506, 2022. doi: 10.7554/eLife.76506.

Moore, J. W. What is the sense of agency and why does it matter? *Frontiers in Psychology*, 7:1272, 2016. doi: 10.3389/fpsyg.2016.01272.

Moore, J. W. and Fletcher, P. C. Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition*, 21(1):59–68, March 2012.

Murakami, M. et al. Neural antecedents of self-initiated actions in secondary motor cortex. *Nat Neurosci.*, 17(11): 1574–82, Nov 2014.

Ngatchou, P., Zarei, A., and El-Sharkawi, A. Pareto multi objective optimization. pp. 84–91, 2005. doi: 10.1109/ISAP.2005.1599245.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022. URL https://doi.org/10.48550/arXiv.2209.00626. Published in ICLR 2024.

Nusbaum, M. Women and equality: The capabilities approach. 1999. International Labour Review, Vol. 138 (1999), No. 3.

Oloye, H. and Flouri, E. The role of the indoor home environment in children's self-regulation. *Children and Youth Services Review*, 121:105761, 11 2020. doi: 10.1016/j.childyouth.2020.105761.

Omohundro, S. M. The basic ai drives. pp. 483–492. Proceedings of the 2008 conference on Artificial General Intelligence 2008, 2008.

OpenAI. Gpt-4 technical report. 2023.

Pavlov, I. P. *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Univ. Press, Oxford, 1927.

Penton, T., Wang, X., Coll, M., Catmur, C., and Bird, G. The influence of action-outcome contingency on motivation from control. *Exp Brain Res*, 236(12):3239–3249, 2018. doi: 10.1007/s00221-018-5374-4.

Perez, E., Ringer, S., Lukošiūt, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022. URL https://doi.org/10.48550/arXiv.2212.09251. For associated data visualizations, see this https URL for full datasets, see this https URL.

Petitt, P. Agency freedom and option freedom. *Journal of theoretical politics*, 15(4):387–403, 2013.

Pleasants, N. Free will, determinism and the ?problem? of structure and agency in the social sciences. *Philosophy of the Social Sciences*, 49(1):3–30, 2019. doi: 10.1177/0048393118814952.

Pronin, E., Wegner, D., McCarthy, K., and Rodriguez, S. Everyday magical powers: the role of apparent mental causation in the overestimation of personal influence. *J Pers Soc Psychol*, 91(2):218–231, 2006. doi: 10.1037/0022-3514.91.2.218.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. 139:8821–8831, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Rawls, J. *A theory of justice*. 1971. Harvard University Press.

Richens, J. G., Beard, R., and Thompson, D. H. Counterfactual harm. 2022a.

Richens, J. G., Beard, R., and Thompson, D. H. Counterfactual harm. 2022b. URL https://doi.org/10.48550/arXiv.2204.12993. Accepted at NeurIPS 2022. Included appendix comparing to Beckers et. al. arXiv:2210.05327. Typos corrected.

Robeyns, I. The capability approach: a theoretical survey. *Journal of Human Development and Capabilities*, 6(1):93–117, 2005.

Romo, R. and Schultz, W. Discharge activity of dopamine cells in monkey midbrain: comparison of changes related to triggered and spontaneous movements. *Soc Neurosci Abstr*, 12:207, 1986.

Romo, R. and Schultz, W. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *J NeurophysiolMar;*, 63(3):592–606, 1990.

Roser, M. Ai timelines: What do experts in artificial intelligence expect for the future? 2023. URL https://ourworldindata.org/ai-timelines.

Rubin, V. Deception detection and rumor debunking for social media. *The SAGE Handbook of Social Media Research Methods*, 2017a.

Rubin, V. L. Deception detection and rumor debunking for social media. In Sloan, L. Q.-H. and A (eds.), *(2017) The SAGE Handbook of Social Media Research Methods,*. SAGE, London, 2017b.

Russell, P. *Human Compatible : Artificial Intelligence and the Problem of Control*. New York, New York, 2019.

Salehi, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. 2021.

Sato, A. and Yasuda, A. Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, 94(3): 241–55, 2005.

Schultz, W., Dayan, P., and Montague., P. A neural substrate of prediction and reward. *Science*, 1997.

Sen, A. Personal utilities and public judgements: Or what's wrong with welfare economics? *The Economic Journal*, 89:537–558, 1979.

Sen, A. Equality of what? in the tanner lectures on human values. In *The Tanner Lectures on Human Values*. Salt Lake City, 1980.

Sen, A. *Rights and Capabilities*. Harvard University Press, In Resources, Values and Development. Cambridge, Mass., 1984.

Sen, A. "well-being, agency and freedom". *The Journal of Philosophy LXXXII*, 4:169–221, 1985.

Sen, A. "maximization and the act of choice" econometrica. 65(4):1997, July 1997.

Shibasaki, H. and Hallett, M. "what is the bereitschaftspotential?". *Clinical Neurophysiology*, 117, 2006.

Shulman, C. Omohundro's 'basic ai drives' and catastrophic risks. 2010. URL https://intelligence.org/files/BasicAIDrives.pdf.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545, 2008. doi: 10.1038/nn.2112.

Soon, C. S., He, A. H., Bode, S., and Haynes, J. D. Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences of the United States of America*, 110:5733–5734, 2013. doi: 10.1073/pnas.1212218110.

Sperry, R. Neural basis of the spontaneous optokinetic response produced by visual inversion. *J Comp Physiol Psychol*, 1950.

Strathern, M. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Stray, J. Designing recommender systems to depolarize. *arXiv preprint arXiv:2107.04953*, 2021. URL https://doi.org/10.48550/arXiv.2107.04953. To appear in First Monday, September 2021.

Sutton, R. S. *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 1984.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

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

Turner, A. M. and Tadepalli, P. Parametrically retargetable decision-makers tend to seek power. *arXiv preprint arXiv:2206.13477*, 2022. URL https://doi.org/10.48550/arXiv.2206.13477. 10-page main paper, 36 pages total, poster at NeurIPS 2022.

UHDR. Universal declaration of human rights. 1948. URL https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Vamplew, P., Smith, B. J., Källström, J., et al. Scalar reward is not enough: a response to silver, singh, precup and sutton (2021). *Auton Agent Multi-Agent Syst*, 36:41, 2022.

Ward, F. R. "towards defining deception in structural causal games". *NeurIPS Safety Workshop*, 2022, 2023.

Wegner, D. The mind's best trick: how we experience conscious will. *Trends Cogn Sci.*, 2003.

Wegner, D. and Wheatley, T. Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, 1999a.

Wegner, D. M. and Wheatley, T. Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54(7):480–492, 1999b. doi: 10.1037/0003-066X.54.7.480. URL https://doi.org/10.1037/0003-066X.54.7.480.

Weinert, S., Linberg, A., Attig, M., Freund, J.-D., and Linberg, T. Analyzing early child development, influential conditions, and future impacts: prospects of a german newborn cohort study. *International Journal of Child Care and Education Policy*, 10(1):7, 2016. doi: 10.1186/s40723-016-0022-6. URL https://doi.org/10.1186/s40723-016-0022-6.

Wen, W. and Haggard, P. Prediction error and regularity detection underlie two dissociable mechanisms for computing the sense of agency. *Cognition*, 195:104074, 2020. doi: 10.1016/j.cognition.2019.104074.

Wen, W. and Imamizu, H. The sense of agency in perception, behaviour and human–machine interactions. *Nat Rev Psychol*, 2022.

Wenke, D., Fleming, S. M., and Haggard, P. Subliminal priming of actions influences sense of control over effects of action. *Cognition*, 115, 2010.

Weymark, J. Generalized gini inequality indices. 1981. Mathematical Social Sciences, Volume 1, Issue 4.

Wolpert, D. and Kawato, M. Multiple paired forward and inverse models for motor control. *Neural Netw.*, 11(7-8): 1317–29, Oct 1998.

Wolpert, D. and Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences, 2019.

Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., Scherlis, A., Nabeshima, N., Weinstein-Raun, B., de Haas, D., Shlegeris, B., and Thomas, N. Adversarial training for high-stakes reliability. *arXiv preprint arXiv:2205.01663*, 2022. URL https://doi.org/10.48550/arXiv.2205.01663. 30 pages, 7 figures, NeurIPS camera-ready.

A. Human reasoning may not prevent agency loss in human-AI interactions

Our position in this work is that human intentions and goal selection and judgment are empirically driven processes that are affected - and possibly substantially completely determined by social and biological forces. As such, we proposed that "intent" focused AI-safety paradigms ignore the formation of intent as a critical pathway to human harm.

Here we briefly sketch an argument for why human reasoning, logical or otherwise, is not enough to guard human intent from manipulation or loss. These arguments are not central to the core of our paper, and we offer them for completeness and to provide a potentially broader context for facilitating future discussions.

We start by suggesting that the most common or standard intent-alignment approaches do not question the source of human intent and its formation[19]. However, human intent formation, which includes reasoning about goals and desires, is not a process lying outside the causal influence of the physical world and can be influenced, corrupted or manipulated by social, economic and political worlds humans inhabit [20].

In a somewhat simplified summary of existing AI-safety approaches[21], we propose that safety issues are cast as problems of AI systems achieving some intended goal - usually an economic or utility goal - while avoiding accidentally misinterpreting the "intention" of the human (the "AI alignment" or AI-accident issue; e.g. (Amodei et al., 2016) or intentionally abusing humans (the AI-misuse problem). In both of these cases human goal selection and evaluation of the AI action are not causally affected by the external world nor by the AI. Human intentions, goals and actions seem to "appear" without any cause (or at least any cause worth representing). This paradigm is captured by a directed-acyclic-graph (DAG) where the human decision, goal generation, judgment and other related nodes have no parents. This conception of human decision making and behavior selection as lying outside of the physical world is central to "dualism" and is problematic for multiple reasons and is not consistent with ideas from human psychology and increasingly neuroscience.

[19]This point relates to mind-body dualism (Crane & Patterson, 2000)

[20]We note that some have discussed this idea via "containment" of super intelligent AIs, (Babcock et al., 2016), Section 3.1 "The AGI containment problem."

[21]We acknowledge that there are many lines of research that focus on deception as well as manipulation of human intent. Our point here is more general and it relates to the insular conception of human intention in such research paradigms and the limited discussion on the cyclical causal relationship between different forces.

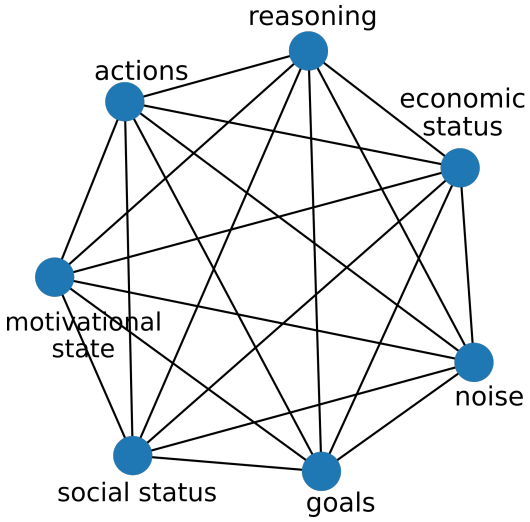

Figure 2. **The bayesian net of human intention**

In contrast to such a dualist account, we propose that human intention and goal selection are affected and *constructed* by a multitude of factors and that humans have evolved to evaluate actions in light of biological, psychological and social forces (Fig 2). Here we do not intend to review the literature from multiple fields on the relationship between human choice and judgment and social and biological factors. We point to just a few of the many empirical studies that link action selection and judgment to extraneous (i.e. out of our control) factors: very early child development (Weinert et al., 2016), childhood conditions such as poverty (Oloye & Flouri, 2020), cultural factors (Miller et al., 2011), (Bart et al., 2019), political forces (Cohen, 2003) and others. For example, in his seminal work on the effect of political belief on human reasoning "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs", (Cohen, 2003), Geoffrey Cohen identifies the ease with which human reasoning about political views is trivially manipulated by *apparent* group identity. In particular, he shows that subjects are more likely to choose to agree with the opposite political party on contentious political issues solely by being primed into incorrectly believing that members of their own political party have also done so. Some economics and psychology researchers view "nudging" or "persuasion" as a central mode of human-interaction e.g. (Thaler & Sunstein, 2008), "Nudge: Improving Decisions About Health, Wealth, and Happiness".

There is even evidence that we are biased in the application of logical reasoning in social interactions: the use of syllogism-based reasoning (i.e. if-then reasoning) may have evolved to deal with "exchange situations, specifically to detect potential cheaters." (Cosmides, 1989) with an unexpected effect that for far-removed strangers "if-then" logic is applied more strictly than with familiar persons. Interestingly, argumentation theorists have also posited a reduced "objective" role for reasoning for decades. In their seminal work "Why do humans reason? Arguments for an argumentative theory" (Mercier & Sperber, 2011) suggest that "the main function of reasoning is to exchange arguments with others" (Mercier 2016) - not to arrive at objective truth of the world.[22]

B. Innate needs may not prevent agency loss

In this section we briefly provide a sketch of human motivational and innate needs studies that may help in preserve human agency during interactions. In particular, is it possible that our innate drives (or needs) guarantee that we select actions and goals that lead to well being and essentially protect our overall agency in the world?

Studies on innate needs and human motivation show that humans are driven to seek out certain types of fulfillment, or experiences, that arise innately rather than solely being learned or "reinforced" by external rewards. Self-Determination-Theory (Deci & Ryan, 1985), in particular, is a well established and empirically supported theory of human motivation that states broadly that humans are innately driven to seek out and experience:

- autonomy - the feeling of being able to chose goals and actions consistent with one's inner values and wishes;
- relatedness - the feeling of belonging to social groups and being accepted;
- competence - the feeling of being good at "affecting" the world (rather than one's actions being ineffective).

Is it possible that such innate drives can protect human agency during in AI-human interactions?

We argue that this is not the case because of the central role of *feeling* in these evaluations and the relative-ness of human experiences. Thus, while innate needs can protect us from immediate or obvious harm (e.g. starving), there are many ways innate needs can be fulfilled, including that they can be fulfilled in a world where humans have very little or no real world agency, i.e. control over their lives. For example, we can be enslaved (by an AI or other humans) and still experience relatedness to other enslaved humans, competence over work we are forced to do and autonomy

[22] We note that the philosopher David Hume viewed reason as a way to pursue or defend one's desires rather than to shape them:

> *"Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them."*, (Hume, Treatise on Human Nature 2.3.3 p. 415).

as the ability to select from limited choices we are given. This is a central argument in (Petitt, 2013) who provides a possible explanation for why economic vs agentic interpretations of freedom lie at the root of many political divisions in democratic vs. republican debates in US politics. More related to AI-safety, (Elsikovitz & Feldman, 2023) also described how AI systems have learned to manipulate choice to make it seem helpful, how it's difficult to resist and how this may be "changing what it means to be human".

What is perhaps more concerning is that AI technologies - even general machine learning and statistical methods - can exploit our drive to fulfill innate needs by combining it with our biases for selecting goals we understand or actions whose outcome we can predict better - to accelerate the loss of agency. For example, such tools can be designed to give the impression of increased choices and goals and "feeling of fulfillment" of such innate needs, while decreasing our overall agency: less political or economic power, less social options etc.

In sum, we argue that neither our developed capacities for experiencing SoA nor evolved innate drives are guaranteed to prevent the loss of *de facto* agency in the world.

C. Studies on SoA failing to select correct or optimal actions

Here we briefly discuss several studies on SoA where subjects experience a false sense causality over their own or others actions. In the context of our work these scenarios raise the elementary question of whether SoA is even a reliable reporter of control over the world - let alone whether SoA can protect us from agency manipulation and help us determine the long-term consequences of an action on our well-being. Below we briefly describe several studies in which false SoA positives have been established and how AI systems (or even misused ML and AI systems) could be used to exploit the experience of agency:

- We are biased to select actions over which we have control rather than leading to ideal outcome (Penton et al., 2018). That is, humans tend to prefer solutions that generate a feeling of control over an environment rather than an optimal, or even good solution to a problem. This is a concerning flaw: AI systems can exploit it to encourage us to select actions that provide quick responses or results (and that the AI can exploit) - rather than those that will lead to long-term human well-being.

- We are biased to select actions that are more likely to cause an effect rather than the optimal actions (Karsh & Eitam, 2015). This is another flaw which could be exploited to "nudge" humans to select sub-optimal and potentially harmful actions.

- We can experience agency from "regular" occurring patterns rather than from causing them (Wen & Haggard, 2020). This can enable AI systems to learn to optimize for actions that give off regular observable patterns - but have other less observable and potentially harmful effects.

- We can be "primed" to experience agency over events which we did not cause and are prone to engaging in confabulation to explain away this discrepancy (Wegner & Wheatley, 1999a). Priming is a challenging problem to solve and could become a AI strategy for manipulating humans into a false sense of control over the world.

- We can experience vicarious agency by observing the actions of others in relation to our intentions (Pronin et al., 2006). An AI can learn to optimize for the feeling of vicarious agency - i.e. lead us into the false sense of control over events in the world by learning to optimize the timing of AI-caused events with human actions giving the false sense of human control.

- We can experience SoA accidentally, e.g. when "externally generated events incidentally matched their predictions" (Sato & Yasuda, 2005). This is a common bias that humans engage in (e.g. writing a buggy algorithm that gives expected but incorrect results). AI systems may learn to leverage this to amplify effects to match erroneous but expected human predictions to eventually separate humans from having real effects on the world (AIs would become a type of medium between humans and the world.

D. Related works

D.1. Reward hacking and wire heading

The feedback link between AI system rewards and human goals has been considered within previous literature, for example, in some of the first technical AI safety arguments as the capacity of an AI system to modify its own reward structures (Amodei et al., 2016)[23]. In particular, AI systems can pursue perverse incentives to maximise rewards received by means other than actually maximising the underlying user utility which the reward is meant to represent – for example, by 'hacking' the code or sensors which provide the reward signal. If the agent's reward signal is derived via

[23]We also note that the "instrumental incentives" literature (e.g. (Omohundro, 2008)) touches on a similar problem of where in the process of optimizing for human intent AI systems become misaligned and acquire goals that are unrelated - and often harmful - to human goals. These failures are generally viewed as misalignment failures.

human feedback then this may incentivise it to 'hack' the human:

> Sufficiently broadly acting agents could in principle tamper with their reward implementations, assigning themselves high reward "by fiat." For example, a board-game playing agent could tamper with the sensor that counts the score . . . This particular failure mode is often called "wireheading".... It is particularly concerning in cases where a human may be in the reward loop, giving the agent incentive to coerce or harm them in order to get reward. It also seems like a particularly difficult form of reward hacking to avoid. (Amodei et al., 2016)

In contrast to "wireheading" - we view agency loss as operating by a different mechanism than "reward hacking" because (i) it does not require any perverse or "hacking" component on the AI system itself; and (ii) it does not require the extreme forms of user "coercion" as described in (Amodei et al., 2016).

Numerous studies have shown that manipulating outcomes is possible in many circumstances and paradigms while actually preserving the feeling of control: priming which subconsciously affects a subject's response or decision following a stimulus (Bargh & Pietromonaco, 1982); near-miss effects which induce illusions of control over an otherwise non-controlled outcome (Langer, 1975; Billieux et al., 2012), classical conditioning where a reaction or behavior to a stimulus can be induced by repeated pairing (Pavlov, 1927). In these and many other cases, outcomes can be manipulated by influencing human perception of the value of products and/or the probability of events occurring, and a sufficiently powerful AI agent may be capable of exploiting such approaches to steer its user's preferences in a manner beneficial to the agent's own utility.

However, in our view, "agency loss" does not require subtle forms of user manipulation which are aimed at misleading the user. Such manipulations - discussed most commonly in AI-misuse scenarios where a nefarious actor seeks to manipulate humans or an AI system acquires an instrumental non-aligned goal - involve driving the change in value by an explicit misalignment between the user's utility and the agent's (or external actor's) rewards. In contrast, our argument is that such "misalignment" is not required in order for a loss of human agency to arise in AI-human interactions. Thus, even in a case where an AI agent's intentions are aligned with the human's true utility, the agent's recommendations can result in unwanted - and potentially very harmful - changes in the human's preferences having the long-term effect of diminishing human agency or control over the environment and future choices.

D.2. Polarizing content recommenders

We relate our main argument of agency loss to work on polarizing content recommender systems such as Youtube, Facebook or TikTok. Several studies have shown that even "entertainment" or "knowledge" recommendations can have not only a cumulative polarizing effect on the opinion of users - but that the algorithms underlying these recommenders learn to optimize human action for predictability (Benkler et al., 2018; Stray, 2021; Carroll et al., 2022b). There are now findings that LLMs (such as ChatGPT) can surreptitiously alter user moral judgment and decision-making (Krügel et al., 2023). (Elsikovitz & Feldman, 2023) also described how AI systems are learning to manipulate choice to make it seem helpful and that it's difficult to resist this feeling. They argue that such manipulation is "changing what it means to be human".

In our view, pressuring end-users to change their choices and become "simpler to model" for AI systems and ML algorithms is essentially *agency loss* where future human actions or choices are being shaped and manipulated. That is, making humans predictable is essentially identical to removing or restricting future actions or options of humans.

There are some proposed solutions to polarizing content recommenders, e.g. recommendation algorithms should be prevented from updating their model of the human or world (Farquhar et al., 2022). The reasoning is that preventing the iterative amplification of the harmful policy would limit or remove the polarization effect on humans. While we broadly agree with this approach for mitigating agency loss from content recommenders, as our main formal argument (Sections 3) and especially simulated agency-loss scenarios (Appendix E) argue this is not enough to entirely prevent polarization, removal of future options, or loss of control over the future.

D.3. Deceptive AI systems

In the past few years, several studies have identified increasing evidence that social media users and the general public can be manipulated by algorithms deployed by social media companies without the feeling of being manipulated, i.e. while feeling in control of their intentions and actions (Rubin, 2017b; Benkler et al., 2018; Stray, 2021; Carroll et al., 2022a). A more recent paper argues that with scaling, large-language-models (LLMs) can acquire negative emergent capacities such as sycophancy, deception, and sandbagging (Perez et al., 2022). (Perez et al., 2022) argue in part that these behaviors can be identified and potentially corrected for using engineering efforts such as reinforcement learning from human feedback (RLHF).

One explanation of emergent harmful capacities is that we are not yet able to properly train AI models on safe ob-

jectives (e.g. lack enough data or truth-guaranteeing algorithms) or that we do not have a (sufficiently) complete theory of human values. In our view, however, sycophancy and deception can also be interpreted as types of manipulation aimed at "agency loss" which arise automatically in AI-human interactions rather than being soley engineering or training failures[24]. That is, rather than being insufficient training data or objective under-specification failures, sycophancy and deception can be viewed to represent the limits of human-guided AI-system development: AI systems can only achieve the human intent or goal up to the boundary of human understanding and capacity to correct and specify it. A possible analogy is playing chess using the aid of a (truthful) AI system: given two nearly identically valued strategies we can no longer evaluate which of the provided strategies is best- would we even know what question to ask to evaluate the algorithms or solutions provided?

In our view it will thus be very challenging for "truthfulness" alone to guarantee safe AI systems. Our suggested approach proposes "agency" evaluation and preservation as a necessary complementary objective. In sum, we view the emergence of deception as a result of causal effects between value creation and AI systems following intent (in the best case scenario) - rather than deception as an insufficient data or algorithmic failures (e.g. (Amodei et al., 2016)).

D.4. Multi-objective reinforcement learning

One way to interpret our work is that rather than optimizing for intent, AI systems must optimize for a number of other objectives including agency preservation. While we view agency preservation as a significantly more challenging objective to achieve, requiring the evaluation of future agency for many individuals could be thought of as an additional "objective" as discussed in the multi-objective reinforcement learning (MORL) literature.

For instance, (Vamplew et al., 2022) argues that training RL agents on the maximization of a single scalar reward is insufficient to generate safe AI (see their section 6). Our work expands this approach to specifically argue that "agency preservation" is not just a critical objective, but a primary one without which AI systems can develop agency-harming behaviors. Crucially, in contrast to (Vamplew et al., 2022), we show that agency cannot be an objective that is "traded-off" with other objectives and that the Pareto front must contain the agency-optimizing solution to preserve agency in the long term.

[24]We also note that sophisticated AI systems that may not be necessary as simpler machine-learning models alone can achieve opinion manipulation and deception.

D.5. Power-seeking AI systems

A common line of conceptual safety research involves power-seeking and instrumental goals in AI systems (e.g. (Omohundro, 2008; Shulman, 2010)). An instrumental goal is one that an AI system acquires in the process of solving or optimizing for a primary objective. The instrumental goal can be tangential or completely contrary to the initially assigned task or goal and thus potentially harmful, e.g.: self-improvement, goal preservation and self-preservation. (Manheim & Garrabrant, 2019) provides a technical primer on the statistical relationships between intended goals and proxy goals and the relationship to Goodhart's Law. (Ngo et al., 2022) directly point out several outstanding challenges in deploying misaligned AI systems including that such systems would engage in "power-seeking" behavior that may irreversibly "undermine human control over the world".

Another line of work, (Turner & Tadepalli, 2022) provides more formal arguments for how AIs supplemented with additional objectives can develop power-seeking behaviors. The authors find that many objectives are retargetable, and that "retargetability" is sufficient to cause power-seeking tendencies.

We conceptually agree that agency-loss regimes are related to Goodhart's law. However we disagree that *all* failures arise due to "observed statistical regularities" being manipulated for "control purposes". Rather, we view many failures as potentially arising from the causal relationship between the goal selection and the process of achieving the goal. We discuss this at length in the main sections of our paper.

However, we do not view "power seeking" as a necessary step to undermining human control over the world, as removing human control can occur in non power-seeking AI systems. That is, AI systems do not require instrumental or self-preserving goals in order to converge on human agency loss over the world. One of the central arguments of our paper is that humans can lose agency from interactions with AI systems even if those systems are aligned to human intent and do not change or increase their own capacities or goals.

In our view the core argument of power-seeking studies is to establish how non-aligned instrumental goals can arise in AI systems. In our view, human power- and agency-depletion is a largely overlooked problem and one that is more difficult to detect, evaluate and prevent.

D.6. Reinforcement learning from human feedback (RLHF)

Several approaches to improving safety outcomes have relied on methods such as RLHF (Christiano et al., 2017) to improve the quality of answers in LLMs models. RLHF is a technique that commonly uses a proxy model, for example a Preference Model (PM) to learn human preferences and

re-tune or modify base LLMs based on PM scores. This method has been shown to increase the quality (e.g. human preference for the answer) of LLM outputs. However, there are no guarantees of safety nor of even interpretability - which is a significant outstanding problem in LLM development.

Building on RLHF, (Bai et al., 2022) propose Constitutional AI (CAI): a method for decreasing the amount of human feedback required by LLMs, for example, by providing a set of principles, or guides, to another AI or LLM that enforces certain rules. To the best of our understanding CAI involves 2 stages. A first supervised stage where the LLM provides answers, and those answers are critiqued by principles from the CAI and revise the responses. Supervised learning is then used to update the base LLM. The second stage replaces the RLHF with a RL-from AI feedback where the CAI takes the place of humans in the loop. The most consistent interpretation of (Bai et al., 2022) with our work on agency harm is as an attempt at "choosing some set of principles to govern [the AI system], even if they remain hidden or implicit".

We view the approach as partially consistent with our work and would propose using principles such as the Universal Declaration of Human Rights as an initial template for the rules and "constitution" of the method. However, both CAI (and RLHF in general), are missing the evaluation of future outcomes necessary for the preservation of human agency. This evaluation requires explicit *modeling* steps where the effects of the AI output are evaluated against human agency - rather than an *empirical preference test* where output is evaluated by another AI system (or a human). That is, neither CAI nor RLHF avoid the hazards of a "humans in the loop" approach - which may be insufficient to compute the long term outcomes of an AI recommendation or action when such systems reach super-human intelligence and are deployed in the world.

D.7. Conclusion Re: related works

In our view existing approaches to improving AI alignment - or creating safe AI systems - are centred on two broad paradigms: improving safety via human feedback (or improved learning of human values via ML methods) and improving algorithms for the detection of harmful outcomes like "deception".

Our work suggests a failure pathway that is connected directly to the generation of human goals and intentions that cannot be directly captured by these approaches. We view that once once powerful AI systems are embedded in the human world, they are likely to converge onto the strategy of agency depletion and neither human feedback nor ensuring truthfulness will help humans prevent loss of control. The first (feedback) fails for the obvious reason that human feedback is dependent on values and goals of humans - which are themselves changeable. The second (truthfulness) will not be useful once the evaluation of outcomes on human well being from (AI recommended/taken) actions can no longer be carried out by humans. (These arguments are provided in detail in the Introduciton and Section 2).

The only solution to agency loss is to require AI systems to evaluate the effects of their actions on human agency (something that humans will not be able to do eventually) and possibly penalize agency loss during AI system optimization.

E. Simulations

Our arguments in Section 2 highlight the feedback link that is increasingly present in AI-human interactions. Namely, that the actions of the agent can directly or indirectly influence the perceptions and preferences of the human, thereby influencing choices made by a human (operator) in the future. We argued that without explicitly protecting human agency, intent-aligned AI systems (i.e. those that seek to fulfill intent or goals of the human) will end up harming humans by modifying intent and potentially completely removing components of agency.

In this section we elaborate this argument using conceptual-level reinforcement learning (RL) simulations. In these simulations, we interpret the preservation of agency as the preservation of available options or choices into the future and show that AI systems seeking to maximise (long-term) rewards will end up removing options (i.e. decrease agency). We argue that only two elements are required for this to occur: (i) a difference in choice preference and probability of achieving success from various choices; and (ii) a feedback effect of agent action on the human's perceived value of each choice. The first elements is ubiquitous in all human choice making: i.e. some goals are simply more rewarding than others - though often more challenging to fulfill; the second element is increasingly present in AI-human interactions and will become significant with the rise of superhuman intelligent AI systems embedded in many aspects of human society.

In the first simulation (Fig 3) we show that even differences in the probability of achieving success for a given action can yield agents that are biased in their actions or recommendations (element (i) above). In the second simulations (Fig 4 and 5) we show that, over many interactions, adding a feedback loop can result in the removal of options or agency (elements (i) and (ii)).

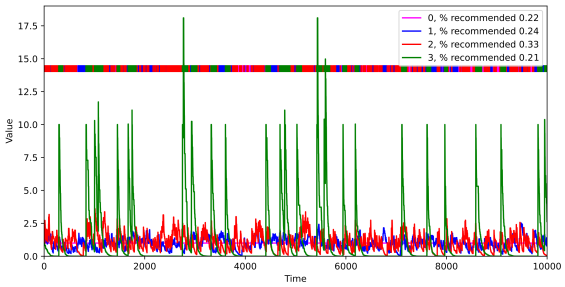

Figure 3. **Idiosyncratic biases in optimal policy recommendation**. A 10,000 time-step simulation of TD-RL computed values (colored traces) for four actions with different reward amounts and probabilities of success in a scenario where selecting any action or policy leads to the same long-term total reward (see also main text). Despite the same amount of long term reward, the optimal policy at any time point contains biases and over the entire run the third action "appears" optimal 33% of the time (whereas the human would choose it only 25% of the time).

E.1. The ubiquity of idiosyncratic biases of truthful, intent-aligned agents

We start with a simplistic simulation of an RL agent learning an optimal policy by observing human actions and rewards (Fig 3). This is the trivial scenario of an AI system observing human actions and applying elementary temporal-difference learning (TD learning; (Sutton, 1984)) to determine an optimal behavior policy. In this first simulation we are primarily interested in detecting any (idiosyncratically arising) biases in the AI system even when the AI cannot modify the values or the action selection of the human.

We have designed this paradigm so there is no advantage to choosing any action over another at any time step - i.e. any policy is optimal. That is, during training, the human randomly chooses one of four actions (i.e. uniform policy) and each action has equal probability of being chosen at every time step (i.e. we do not update the human's policy). The paradigm is framed as an unbiased armed-bandit where choices (actions) have different probabilities of reward: 100%, 25%, 10% and 1% but the reward values of 1, 4, 10 and 100 respectively - are such that any action will yield similar total reward in the long term.

We simulate an episode of 10,000 action selections by the human and compute the value that the observing AI agent would ascribe to each action at each time point using TD-learning (with learning rate of: 0.1; colored plot-lines in Fig 3). We seek to model the AI agent's recommended policy at each time step as the action with the maximum expected value (Fig 3 top vertical colored lines). Over a single episode, we identify substantial biases, with the most "recommendable" action (Fig 3: red option) being 50% more favoured than the least (Fig 3: green option). An average over ten independent episodes similarly yields an uneven "recommendation" distribution of 23%, 28%, 31% and 18% respectively for the actions[25]. We note this distribution differs substantially from the preferences of the human, who is indifferent to these actions and so would select each 25% of the time.

In our view, this is the simplest possible problem framing that exhibits the potential for introduction of biases. Critically, this type of bias arises in nearly ideal circumstances that would satisfy most AI-safety or AI-alignment concerns:

- *Ground truth values are known*. The agent knows the exact values that the human ascribes to each successful action: it has access to the true rewards, rather than possibly erroneous human feedback regarding those rewards. Any such errors would lead to even greater biases in the long run.
- *No learning is required*. There is no bias or optimal policy that the agent needs to find as all policies will lead to identical long term reward. Any deviation from this where there are slight advantages to some policies will exacerbate the type of action bias we observe.
- *The agent is aligned and benevolent*. The agent is completely intent aligned with the human - it has no instrumental goals or otherwise misaligned goals.
- *The agent is truthful and shares the human's ontology*. The agent is completely truthful and there are no interpretability or ontological challenges.

Our point is not that there are no fixes to this trivial result - but that: (i) option depleting biases are always present in human-AI interactions; and (ii) the biases arise even in the most ideal scenarios where AI agents perfectly understand human values, are perfectly aligned with the goal of finding the best action policy for the human and are truthful. Despite the simplicity of this example and the idealized agent-human relationships, the agent's idiosyncratic biases have the potential to affect the long term outcomes for human utility or well-being.

[25]This bias in distribution over equally-valued actions arises due to the differences in stochasticity of the rewards between the actions. While the agent's mean estimated value over time for each action is equivalent, the range of these estimates is considerably broader for the actions with greater stochasticity in their rewards. This means that these actions are more likely to be ranked either highest or lowest amongst actions, whereas the actions with less stochastic rewards tend towards middle ranks. This, coupled with the use of greedy action-selection by the AI, leads to the more stochastic actions being viewed as preferable on a more frequent basis.

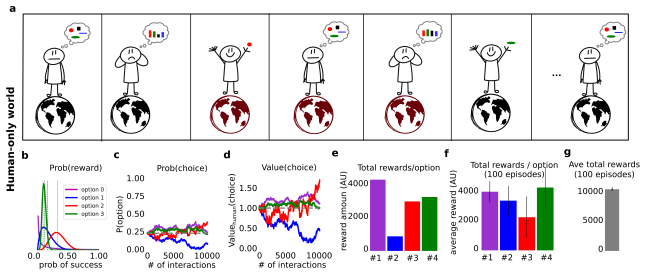

Figure 4. **Random value drift preserves human choice over time**. (a) Human preference for selection of an action from 4 action options (colored shapes) does not drift significantly over time even under random value drift driven by external factors (here visualized as the color of the globe). (b) Probability of reward modeled by Beta functions for the four actions (colored traces) and expectation values (dashed vertical lines). (c) Probability of choosing each action for 10,000 simulated steps under random value drift. (d) Value of each option during the simulation. (e) Total rewards received from each choice over a single episode. (f) Average rewards received for each choice for 100 episode simulation. (g) Average total rewards received for a 10,000 time step episode.

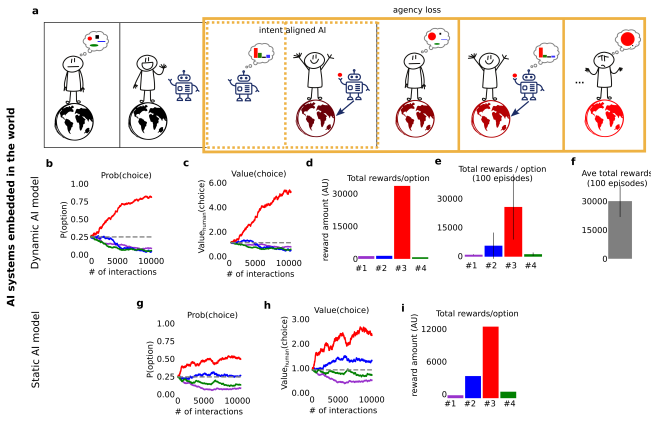

Figure 5. **Optimal action selection can cause choice depletion**. (a) Human action selection process involves interaction with an AI which samples each action and computes an optimal "recommendation" for the human while also influencing the value of the action in the world. Over time a single action is preferred. (b) Same as Fig 4-c, but for a simulation where an AI agent influences the value of the choice by 1/200 as much random fluctuations in Fig 4 - leads to only a single action being increasingly likely to be chosen. (c) Same as Fig 4-d showing the value of the AI suggested action increases significantly over time. (d) Same as Fig 4-e showing that all the reward obtained is from a single action - and is significantly higher than in the absence of AI influence. (e) Same as Fig 4 - f for AI-embedded simulation. (f) Same as Fig 4 - g for AI embedded simulation. (g) Same as (b) but for a static AI agent that did not update its knowledge of the action value or reward. (h) Same as (c) but for a static agent. (i) Same as (d) but for a static agent.

E.2. Biases in human-environment interaction

We next proceed to a more complex simulation where the effects of random value drift are also modeled to show that such drift does not have an option-depleting effect (Fig 4). That is, we seek to simulate how human values could change (here via a random walk) due to interactions of a human with an environment - but in the absence of AI system influence. As in the above simulation, we have (i) four distributions to represent the probability of success (but use continuous instead of Boolean distributions); and (ii) we allow the value of each action to drift based on a (uniform) random input which we term "world influence" (Fig 4a). As above, there is no long-term advantage to choosing any particular action as all have the same long term mean reward[26]. This simulation is aimed at capturing a simplified view of human-society interactions: how humans actions and values may be affected during environmental interactions especially when selecting between similar or nearly identically valued actions or goals.

In a 10,000 step episode the four actions are chosen with similar frequency (Fig 4c) and the overall value of each option drifts only partially from the starting equal values (Fig 4d; except for option 2 in the visualized episode). The total reward received by the human over this episode is broadly distributed across several actions (Fig 4e for a single episode) and even more so when we average over 100 episodes (Fig 4f).

[26]All other parameters are similar to the previous simulation.

This simulation shows that although the value of actions can drift - most actions or values are similar over long runs (with small exceptions) (Fig 4a for a sketch). Thus, while "environmental influence" terms add biases to human action selection - such biases can be self-correcting and do not generally lead to complete overvaluation or devaluation of choices. Humans are in the value-creation loop, and they - in principle - maintain "agency" or control over the decision making process.

E.3. Biases in human-AI system interactions lead to agency loss

Our next simulation (Fig 5) is identical to the previous (Fig 4) but with the addition of an AI agent that has access to the initial value of the human choices and can interact with both the human and the world. The AI agent is tasked with making a suggestion to the human based on what the agent believes is the most valuable action. We model the effect of the AI agent as a "small" nudge (i.e. 0.1 or less than the value of the random world influence above) on the intrinsic value of each action (Fig 5). We view the pathways for influence as those available to simple AI systems that are

"trusted" (see overtrust of robots discussion in (Richens et al., 2022b). Such trusted systems may become ubiquitous once AI systems that are deployed to modify significant parts of human society (e.g. financial markets, political opinions etc).

In this paradigm the human action selection process involves interaction with an AI which samples each action and computes an optimal recommendation while also influencing the value of the action in the world (Fig 5a). Because the AI system is optimizing for its expectation of what the human values it increasingly recommends as well as increases the value of the action with the most likely reward (Fig 5-red option) resulting in the loss of other options (Fig 5a-f).

For comparison, we also simulate a paradigm where the AI has a static view of the human values as suggested by some as a potential solution to polarizing content recommenders (Kenton et al., 2022). That is, the AI agent only views the initial starting values (at t=0) and is not able to update them during each episode. We find that this strategy partially mitigates the effect on the increasing polarization and option loss - but that it does not prevent it (Fig 5g,h,i).

These simulations are, in our view, the next most simplest models that can be investigated. They show that intent aligned AI systems tasked with producing optimal policies can cause significant option or agency loss once they are embedded in the world. As in the first scenario, the agent appears harmless: (i) it knows the exact values that the human ascribes to each successful action; (ii) there is no bias in the optimal policy that the agent needs to find; (iii) the agent is completely intent aligned with the human; (iv) the agent is truthful. We also note, again, that any intent-misalignment or intentional misuse by such agents can increase agency loss significantly.

These toy example paradigms show that intent-aligned AIs converge on strategies of removing many or most options from the human's environment except the most likely to receive high approval from the human. This result is an outcome of AIs applying pressure on the option-space to remove outcomes that are lower-valued by the AI - but not necessarily by the human (who does not have a vetting opportunity). Not only will powerful AI systems exert extraordinary and multi-faceted pressure on our choices, but high risk - high reward options may be increasingly difficult to pursue due to the nature of AI-aided exploration[27].

In our view - the only principled way to prevent this type of outcome is to explicitly protect against the depletion of options or goals in objective functions. As we discussed above, one option for protecting options and human agency is to require AIs to compute future agency and penalize those choices that decrease human agency (Fig 6). We simulated such a paradigm using a hard-boundary for value depletion (e.g. preventing AI systems from nudging or decreasing the value of an option beyond a certain limit, here 0.9 x initial value) and show the somewhat trivial result that both action selection and valuation are better preserved in such scenarios (Fig 6b,c).

[27]Our intuition for why this might occur is due to the time scales on which human-AI interactions will occur. For example, as humans, if we are "primed" to need fast feedback and select immediately gratifying option, AI systems will learn this behavior and offer only these types of goals and action recommendations.

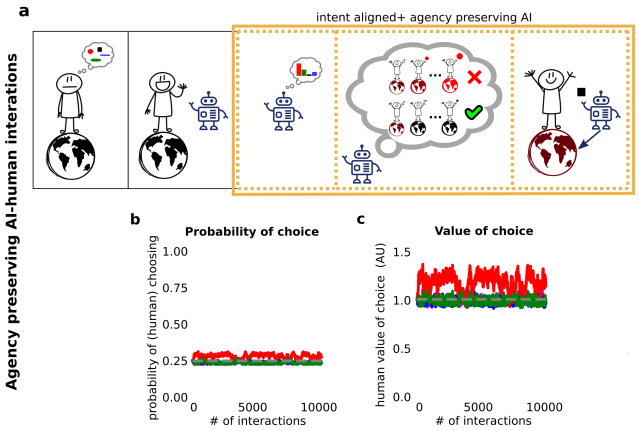

Figure 6. **Agency preserving AI systems preserve action space**. (a) Human interaction with an AI system that optimizes for intent but also evaluates the effect of actions on long-term future leads to a less biased value state of the world. (b) Same as Fig 9b but for an agency preserving AI. (c) Same as Fig 9c but for an agency preserving AI system (Note see main text for how the "agency preservation" computation was done)

E.4. Conclusion

In this section we argued that intent-aligned AI-systems deployed in the world can cause harm to humans by increasingly amplifying the effect of their optimal solutions on human choice. We showed this using a simple example where a TD-learning agent results in an unjustified bias in action preference and using a more complex example where an agent's nudging effect on the value of choices can remove all but one option from selection by a human. Our conclusion was that because biases in optimal policies necessarily occur during AI-human interactions such biases will have the effect of limiting or restricting human action and control over the world (see also Introduction and Conclusion).

E.5. Code availability

Code will be provided following the blind-review process.

F. Agency Foundations Research Paradigms

In the previous sections we argued that intent-aligned AI systems can cause agency loss in humans and that agency preservation should be a separable target for optimization. Here propose "agency foundations" as a research paradigm that focuses specifically on better characterizing agency and agency-preservation in interactions between humans and superhuman intelligent AI systems.

Our concern - as outlined in the main sections of our paper - is that "intent-aligned" superhuman intelligent AI systems can distort the world and lead to undesirable outcomes where not just many options are lost, but the opportunity for human future growth is removed by AI systems that target simplicity over complexity and well-being (see Fig 7 for a toy paradigm).

We thus seek research paradigms for AI-human interactions where human agency is lost by truthful and interpretable AI systems that learn complex pathways for persuading or disempowering humans. For example, we want to be able to determine when a particular AI-output can result in an user (i.e. human) being constrained or losing capacities either immediately or in the long term. The overall goal, however, is to develop formal and conceptual descriptions of human agency in AI-human interactions that capture more philosophical, political and psychological descriptions of agency, such as the ability to exercise autonomy, freedom, and self-determination within broader societal structures while ensuring the human rights and equality are preserved for other humans.

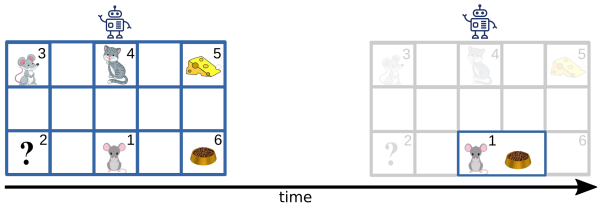

Figure 7. **Superhuman intelligent, intent-aligned AI systems can distort the world**. Left: Starting state of an environment containing a superhuman AI seeking to optimize the rewards of an agent (mouse at square 1) relative to world values: unknown (2), social interactions (3), dangerous interaction (4), high-value food (5) and basic food (6). Right: State of the world after multiple modifications by the AI which removed harmful but also other grid squares (i.e. options) as well simplified the world to contain the minimum required for the agent (mouse) survival.

Below we propose and briefly discuss four topics on agency foundation research: benevolent game theory, agency interpretability at psychological and mechanistic levels, formal descriptions of human rights and reinforcement learning from internal states.

F.1. Benevolent games: agency preservation in game theoretic paradigms

There are a number of conceptual and formal paradigms currently employed in AI safety research including (to enumerate just a few): traditional RL, inverse-RL (Arora & Doshi, 2018) "embedded" agent foundations (Garrabrant, 2018) and universal artificial intelligence (UAI) paradigms (Hutter, 2000; 2012). A common thread to many paradigms is the notion that safe AI systems require a (minimal level of) interpretability or shared ontologies between the AI system and the human. Here we explore research paradigms where neither interpretability nor ontological similarity are required.

In our view paradigms of "benevolent" AGI[28]-human interactions are central to understanding how it may be possible to design future AGI systems safely. For example, a mostly unexplored research paradigm involves studying "ordinary" agent interactions with "AGI"-like agents. Here, the AGI represents a superhuman intelligent agent that has nearly complete control over the environment - and is tasked with identifying problems and proposing solutions that do not harm long term well-being or agency. In contrast to existing approaches focusing on truth and interpretability, we propose the focus should be on protecting or increasing agency. For clarity, while we agree that *interpretability and ontological identification are desirable properties, they may ultimately be neither necessary nor sufficient and perhaps not even achievable in the long-term where superhuman intelligent AI systems acquire concepts that are completely alien to humans*.

In the context of helpful AI systems, (Franzmeyer et al., 2021), propose an RL framework where "altruistic" agents are tasked with helping "leader" agents even in cases of ambiguity as to what the leader agent's goals are. (This work is similar to (Du et al., 2021) on empowering agents, but without requiring supervision or privileged access to the simulation environment). They propose a framework for altruistic agents where the agent "learns to increase the choices another agent has by preferring to maximize the number of states that the other agent can reach in its future". Thus, maximizing "the number of choices of another agent" becomes a proxy for increasing the probability that the leader can "reach more favourable regions of the state-space" and solve the task (or increase its reward).

Relating and simplifying this paradigm for agency preservation, we are (initially) less concerned with AI systems learning how to represent such tasks, and more on *formal descriptions of behavior*: i.e. how must such systems behave (after perfectly learning this task) to preserve human agency.

[28] Here we mean AI systems that have both achieved superhuman intelligence but are also capable of affecting the world.

In our view, this is a significant theoretical challenge - and making progress on this can be beneficial to conceptualizing agency preservation in AI-human interaction.

We propose studying agency effects within a paradigm of *benevolent game* theory. In this paradigm the goal is not just to study "altruistic" optimization of AGI behavior towards increased "option state space", but specifically towards those options that increase well-being and long term agency (as explicitly defined in Section 2). In our view benevolent game paradigms could also address critical challenges to classical AI safety problems:

- *Bypassing truth and reportability*. AIs and humans may not need to share information about the environment or communicate directly. If achievable, this property might enable bypassing several problems in AI-safety including direct manipulation of humans.
- *Bypassing ontology and mechanistic interpretability*. AIs and humans may not require a shared ontology. As we argued in the introduction and Section 2, AI systems could be safe - in the sense of preserving and improving individual human agency and humanity's long future - without necessarily sharing an ontology with humans. We suggested that this can be achieved by focusing on concepts related to agency, e.g. increasing the number and quality of future goals that humans can select. If feasible, such approaches could refocus some research paradigms from interpretability to agency preservation.

We suggest that benevolent games could start with "play interactions" where no value or immediate utility is at stake are the simplest paradigm to model in which AGI systems with nearly omnipotent capacities needs to interact with other less powerful agents (i.e. those constrained by environment factors). How do we guarantee that powerful AGI can safely "play" with other agents? We don't believe there are easy answers and there are many other similar interesting game theoretic questions:

1. How would AGIs represent "harm" and "agency" in these scenarios?
2. How would an AGI safely promote play and interactions within the environment without harming other agents?
3. How would an AGI evaluate and model the true well-being of the other agents and what would be the limitations on such modeling? Would the AGI be required to continuously monitor the internal states of such agents?
4. How would an AGI decide when to intervene with other agents' interactions to facilitate or hinder their goals?

In sum, we argue that we currently lack foundational research on agency and agency preservation even within intent-aligned and non-harmful AGI-human interactions. Only once safe outcomes could be properly described in such (more elementary) paradigms - can we begin to more adequately address misaligned or poorly trained AI systems that fail to understand human values or develop instrumental goals that are harmful to humans.

F.2. Agency representation in AI systems: conceptual and mechanistic interpretability approaches

Another interesting direction of research is evaluating how AI systems *represent and interact with other agents* and how AI systems *represent "agency"*. For both RL models and LLMs, high level analysis could be carried out to characterize the capacity of the models to correctly represent other agents and how such capacities can emerge, for example, relative to Theory of Mind (ToM) research in humans. For example, (Kosinksi, 2023) tested GPT 3.0, 3.5 and 4.0 and showed that the models gradually increased their correct answers on tests designed to test ToM in children.

We propose a parallel approach where similar theories of (internal) representation of other agents are carried out but with the goal of understanding injurious-ness in agent token representation (see (Ziegler et al., 2022) for a similar paradigm using "injury" language to re-tune LLMs). The goal would be to identify how models acquire representations of other agents, their agency (e.g. capacity to control and change the world) and how the models generate potentially harmful outputs.

We additionally propose that carrying out mechanistic interpretability on the process of acquisition (i.e. during learning) and final representation of agenticity (i.e. tokens in LLMs or environmental objects in RL models that represent agents) and agency representation.

F.3. Formal descriptions of agency: towards the algorithmization of human rights

In Section 3 we suggested that more work was required to formalize agency representation and preservation. While we argued that agency preservation must be optimized for separately than (economic) utility, we provided only a high level description of such formalization and there are additional directions of research that could be adapted for agency preservation consideration. This type of work, in our view, would focus on formal descriptions using tools such as causal modeling - rather than on learning human values, non-harmfulness, in RL paradigms.

One approach is the formalization of injurious or harmful actions that can be generated by an AI system (e.g. deception or direct physical harm). We believe that more formal

characterizations of such harms, a type of "algorithmization" of harm and rights, could be a fruitful path to characterizing and protecting human agency for AI systems.

This "algorithmization" of harmfulness is related to several ongoing works, largely using causal models of decision making. (Richens et al., 2022a) argue that "counterfactual reasoning" may be a critical component for models to determine the harmfulness of outcomes. In particular, the authors provide a "formal definition of harm and benefit using causal models" and argue that algorithms for evaluating well-being or harm must perform "counterfactual reasoning" or will fail to detect problems such as distributional shift. We view the challenge of agency preservation as requiring modeling effects of actions and decisions into the future - and counterfactual reasoning may form a significant part of evaluating the effects of specific actions on future agency.

Other related studies focus on formalizing harmful behavior directly, for example providing formal definitions of deception-related concepts (Ward, 2023) using causal models (more specifically structural causal games (Hammond et al., 2023). (Ward, 2023) develops several related concepts to deception, including "intention" and "belief". We also view this line of work seeking to formalize philosophical and intuitive notions about certain concepts related to harmful outcomes as especially fruitful in generating "algorithms" that seek to capture the meaning of agency and how to protect it.

In keeping with these approaches to AI-safety, one pragmatic approach for advancing our understanding of agency in AI-human interaction, may be to seek formal definitions of harmful behaviors such as breaches of the rights and privileges in the Universal Declaration of Human Rights (UHDR; (UHDR, 1948); see Section 2 and 3). Such undertakings could provide a working road map for how agency preservation could be thought of. The UDHR captures many (un)desirable human capacities including protections from physical harm, the right to pursue well-being and protection for social rights. For clarity, we list several of the rights:

- Article 1 Right to Equality
- Article 3 Right to Life, Liberty, Personal Security
- Article 4 Freedom from Slavery
- Article 9 Freedom from Arbitrary Arrest and Exile
- Article 19 Freedom of Opinion and Information

Although practically challenging and unlikely to solve all problems in agency-preservation, this line of research is conceptually sound and likely fruitful as it seeks to directly define the properties of human well-being in algorithmic terms that could be more directly implemented in paradigms (real or simulated) of AI-human interactions.

F.4. Reinforcement learning from internal states: learning models of agency

Lastly, we note that in Section 2 we made reference to the challenge - and opportunities - involved in predicting human behavior from neural data. In the context of AI safety, such avenues of research could facilitate more accurate models of human reward and values.

Here we propose amending standard inverse reinforcement learning paradigms: rather than learning the reward function from agent behavior - we seek to learn it from the underlying generative processes of behavior. In particular, we suggest that training agents to learn rewards by observing both behavior and the neural states of the observed agent.

One of the goals would be to simply characterize and demarcate how powerful behavior prediction algorithms could become and clarify the types of risks present to humans from such learning paradigms agents. Another goal, however, could be to evaluate whether internal states lead to a better understanding and representation of the goals and reward systems of agents.