A Fair Objective for Human-Empowerment-Preserving AI: Desiderata, Design, and Likely Behavioral Consequences

Jobst Heitzig[1] , **Ram Potham**[2]
[1]Potsdam Institute for Climate Impact Research, Potsdam, Germany
[2]Independent
heitzig@pik-potsdam.de, ram.potham@gmail.com

Abstract

This paper explores the idea of promoting well-being and safety in human-AI interactions by forcing AI agents explicitly to empower humans and to manage the power balance between humans and AI agents in a desirable way. Using a principled, partially axiomatic approach based on desirable properties, we design a parametrizable and decomposable objective function for AI systems that represents an inequality- and risk-averse long-term aggregate of human power. It can take into account models of human bounded rationality and social norms, and crucially, considers a wide variety of possible human goals. We prove how certain desiderata enforce particular functional forms and restrict parameter ranges. We exemplify the consequences of softly maximizing this metric in several paradigmatic situations and describe what instrumental sub-goals it will likely imply.

1 Introduction

A duty to empower others, especially those with little power, can be defended on consequentialist [Sen, 2014], deontological [Hill Jr, 2002], and virtue ethical [Nussbaum, 2019] grounds. At the same time, it has been argued that automated AI decision making can gradually *dis*empower humans [Vallor, 2015; Ferdman, 2025; Kulveit *et al.*, 2025].

Accordingly, some papers explored tasking AI agents explicitly with the empowerment of (individual) human users, using the channel capacity $\mathfrak{E}$ between human actions and environmental states, called 'empowerment' in Klyubin *et al.* [2005], as the key metric [Salge and Polani, 2017]. However, that purely information-theoretic metric lacks a clear interpretation in terms of the capability to attain *goals,* which is arguably a key aspect of (em)power(ment) [Lukes, 2026].

Related work Several threads in AI safety address the question of what objective an AI agent should pursue when interacting with humans. Assistance games [Hadfield-Menell *et al.*, 2016] assume the agent is uncertain about the human's utility function and learns it through interaction, but require that a true utility function exists and is identifiable. Impact measures aim to keep agents conservative by penalizing deviation from a baseline; attainable utility preservation [Turner *et al.*, 2020] and reachability analysis [Krakovna *et al.*, 2018] both quantify what an agent's actions foreclose, but do so from the *agent's* perspective rather than measuring *human* capability. The 'empowerment' literature [Klyubin *et al.*, 2005; Salge and Polani, 2017] proposes maximizing the information-theoretic channel capacity between a human's actions and resulting states, but does not account for bounded rationality, goal attainment, or aggregation across several humans. Corrigibility [Potham and Harms, 2025] addresses whether agents permit human correction, typically as a standalone desideratum rather than as an emergent consequence of a unified objective. Our approach connects these threads: like assistance games it is grounded in human agency, like impact measures it cares about preserving options, like 'empowerment' it uses an intrinsic structural metric, and like social choice theory it aggregates across individuals using axiomatic methods. The key difference is that we derive a single, axiomatically motivated objective from which safety properties such as corrigibility, norm-following, and commitment-making emerge rather than being imposed separately.

Contribution Concretely, this paper extends the theoretical foundations of the empowerment-based approach to AI in three ways. First, inspired by the "capability approach" to welfare [Sen, 2014], we develop an *alternative metric* of human power that is directly based on the capability to attain a wide range of possible goals. As our metric is based on some given model of human behavior, it can explicitly and transparently incorporate humans' knowledge about the actions of the AI agent, their expectations about others' behavior, e.g. due to social norms, and their bounded rationality.

Second, using an approach guided by desiderata similar to decision theory, social choice theory, and welfare theory, we develop an objective function for an AI agent interacting with *populations* of humans, based on a fair aggregate of power across humans and along time.

Finally, we analyze the resulting system behavior in a set of stylized situations and argue that the empowerment-based approach gives the AI agent desirable incentives such as communicating well [Reddy *et al.*, 2022], following orders, being corrigible [Potham and Harms, 2025], avoiding irreversible changes in the environment, protecting humans and itself from harm and disempowerment, acting "appropriately" by following relevant social norms [Leibo *et al.*, 2024], and allocating resources fairly and sustainably.

Though based on *possible* human goals, our approach avoids relying on the assumption that the AI system knows or is able to learn individual humans' *actual, current* goals, as it has been argued that human goals are changing and non-identifiable [Cao *et al.*, 2021; Banerjee and Duflo, 2011] and their prediction unavoidably uncertain [Baker *et al.*, 2011].[1]

Our metrics aim to avoid such semantic issues by relating human power to the capability to bring about just *any* possible conditions a human might happen to desire. We are only assuming the AI has (i) a model of how humans behave *conditional* on their goals, and (ii) a *structural* understanding of possibly dynamics, interactions, and transition probabilities. This distinguishes our theory from approaches that form beliefs about humans' *actual* goals, such as "assistance games" [Hadfield-Menell *et al.*, 2016].

The working hypothesis that we would like to start exploring with our work is that using a completely goal-*agnostic* approach already suffices to specify policies for AI systems that make them not only safe but also useful.

For theoretical convenience, we work in a *model-based planning* setting with a stochastic world model [Ha and Schmidhuber, 2018; LeCun, 2022; Zhu *et al.*, 2024; Feng *et al.*, 2025; Bengio *et al.*, 2025]. After developing our human power metric in Section 2, we analyze its likely behavioral consequences in Section 3. A companion paper to this theory paper will focus on scalable estimation of the metrics and empirical validation of the approach. Section 4 concludes.

2 Measuring, Aggregating, and Softly Maximizing Human (Em)Power(ment)

Framework A ***robot*** $\boldsymbol{r}$ (the AI system) interacts with several ***humans*** $\boldsymbol{h} \in \boldsymbol{\mathcal{H}}$. Its ***world model*** is a stochastic game form Γ with states $s \in \mathcal{S}$. Neither r nor h get extrinsic rewards. The humans have many *possible* and potentially *changing* ***goals*** $\boldsymbol{g_h} \in \boldsymbol{\mathcal{G}_h}$, and r pursues an intrinsic objective we are going to design. Action sets $\mathcal{A}_r(s), \mathcal{A}_h(s)$ may be state-dependent. Action profiles are written $a = (a_r, a_{\mathcal{H}}) = (a_r, a_h, a_{-h})$, where $a_{\mathcal{H}} \in \mathcal{A}_{\mathcal{H}} = \prod_{h \in \mathcal{H}} \mathcal{A}_h$. The transition kernel is $\Pr(s'|s, a)$. For simplicity, we assume the game form is finite and acyclic, with terminal state set $\mathcal{S}^\top$.

The robot neither knows nor forms beliefs about the *actual* goals of the humans. Instead, it has a model of human behavior encoded in some given ***goal-conditioned human policy*** $\boldsymbol{\pi_{\mathcal{H}}(s, g)} \in \Delta(\mathcal{A}_{\mathcal{H}})$, where $g \in \mathcal{G} = \prod_{h \in \mathcal{H}} \mathcal{G}_h$ is a goal profile. Although $\pi_{\mathcal{H}}$ does not depend on π_r, human behavior at state s can depend on the robot's *past* behavior because the latter will have influenced what state s the world is in.[2]

[1] Some critics of a goal- or preference-based approach argue for a *values*-based approach instead [Lowe *et al.*, 2025], which however seems to require even stronger assumptions on the AI agent's semantic world understanding. A deep semantic understanding is also required in the 'freedoms'-based conception of AI ethics in London and Heidari [2024], and compiling its required list of 'fundamental capabilities' is difficult [Robeyns, 2006].

[2] We do *not* assume humans play a best response to the robot's *policy* (which is unobserved). Instead, $\pi_{\mathcal{H}}$ contains a model of how humans *adapt* their behavior to *observed past* robot behavior.

Task We want to design an ***objective function*** $\boldsymbol{L(s)}$ that represents *long-term aggregate human power* and fulfills certain desiderata, and a corresponding ***robot policy*** $\boldsymbol{\pi_r(s)}$ that tends to steer the world into states s with larger $L(s)$. We are deliberately not assuming any form of rational choice, optimal control, or welfare theory upfront but rather aim to derive all formulas from first principles. Still, many of the desiderata and resulting equations will look familiar.

The first such principles are aggregation, individualism, and goal-orientedness: We want $L(s_t)$ (where t is time) to be some aggregate of a metric $T(s_{\geq t})$ of *trajectory-specific aggregate human power,* where $s_{\geq t} = (s_t, s_{t+1}, \dots)$ runs over all possible trajectories starting at s_t. Then $T(s_{\geq t})$ shall be some aggregate of a metric $P(s_u)$ of *present aggregate human power* running over all future time points $u \geq t$. Then $P(s)$ shall be some aggregate of a metric $I_h(s)$ of *individual human power* in s, which finally is an aggregate of a metric $C_h^{g_h}(s)$ of individual *goal-attainment capability* in s.[3]

Table 1 gives a preview of the definitions that we will now justify in detail before analysing their consequences.

2.1 Individual-Level Power Metrics

We will not try to capture the full range of subtle aspects of existing notions of human (em)power(ment). Rather, aiming for operationalizability, we focus on aspects we believe a robot can robustly infer from its world model, encoded in state and action sets and transition kernels. So the information we use is rather structural than semantic (meaning-related).

We will see that, as a result, the power metrics are basically functions of the *effective number of goals humans can attain*.

Set of possible goals $\boldsymbol{\mathcal{G}_h}$ We use the simplest possible conceptualization of 'goal': a goal represents what the human would like the world to be like. This can be represented by a proposition about the world state, or equivalently by a (possibly empty) set of states $g_h \subseteq \mathcal{S}$ in which the goal is considered attained. As a very weak form of goal-agnosticism, we desire the following of the set $\mathcal{G}_h$ of h's possible goals:

(G1) Coverage: For all $s \in \mathcal{S}$ there is $g_h \in \mathcal{G}_h$ with $s \in g_h$.

Coverage will also ensure non-negative individual power.

We believe that restricting the model to this very basic type of goal will simplify the derivation of $\mathcal{G}_h$ from learned la-

[3] See the Appendix E for an alternative aggregation order.

$$C_h^{g_h}(s) = 1_{s \in g_h} + \\ + 1_{s \notin g_h \cup \mathcal{S}^\top} \gamma_h \mathop{\mathbb{E}}_{s' \sim s, \pi_h(g_h), \pi_{-h}, \pi_r} C_h^{g_h}(s'), \quad (1)$$

$$I_h(s) = \log_2 \textstyle\sum_{g_h \in \mathcal{G}_h} C_h^{g_h}(s)^\zeta, \quad (2)$$

$$P(s) = -\log_2 \textstyle\sum_{h \in \mathcal{H}} 2^{-\xi I_h(s)}, \quad (3)$$

$$L(s_t) = -\log_2 \mathop{\mathbb{E}}_{s_{\geq t} \sim s_t, \pi} \big(\textstyle\sum_{u \geq t} \gamma_r^{u-t} 2^{-\eta P(s_u)} \big)^\rho, \quad (4)$$

$$\pi_r(s)(a_r) \propto \big(\mathbb{E}_{s' \sim s, a_r, \pi_{\mathcal{H}}} 2^{-L(s')} \big)^{-\beta_r}. \quad (5)$$

Table 1: Derived equations for goal-attainment capability C, individual power I, present aggregate power P, long-term power L as a function of P, and robot policy π_r.

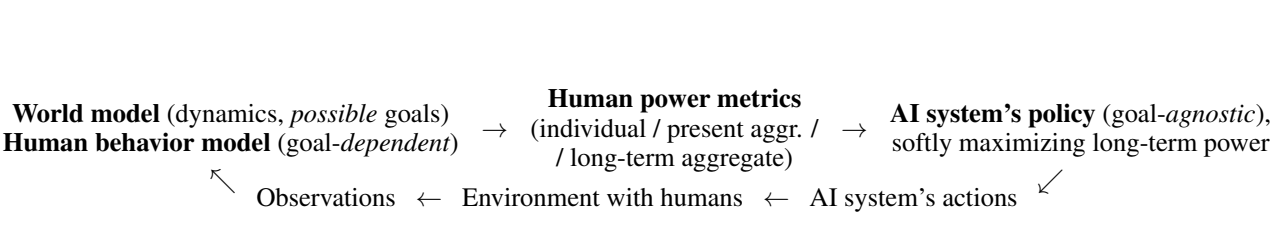

Figure 1: Deriving fair empowerment-preserving AI policies (top), and their feedback on the underlying models (bottom).

tent representations of generic world states and generic human goals as encoded in large language models. As training such models mostly requires statistical data on human goals, it will obviate the need for individual-level data about a particular human's possible goals. This should help mitigating risks arising from misaligned models of human goals.[4]

Goal-attainment capability $C_h^{g_h}(s)$ How should r assess how well a human h would attain a goal g_h? We suggest these axioms should apply across all Γ and $\pi = (\pi_{\mathcal{H}}, \pi_r)$:

(C1) Consequentialism: $C_h^{g_h}(s)$ depends only on the probabilities of trajectories according to π_r, $\pi_h(g_h)$, and π_{-h}, where π_{-h} is averaged over all possible goals of $h' \neq h$.

(C2) Zero Grounding: $C_h^{g_h}(s) = 0$ if and only if g_h is not reached from s with positive probability under π.

(C3) Upper Bound: If $s \in g_h$, $C_h^{g_h}(s) = 1$, otherwise $C_h^{g_h}(s) < 1$ (because goals can become irrelevant over time).

(C4) Recursion: If $s \notin g_h \cup \mathcal{S}^\top$, $C_h^{g_h}(s)$ only depends on the probability distribution of $C_h^{g_h}(s')$ under $s' \sim \Pr(s, \pi)$.

(C5) That dependence is continuous and strictly increasing.

(C6) Independence: If $s_1, s_2 \notin g_h \cup \mathcal{S}^\top$ and $\Pr(s'|s_1, \pi) = \Pr(s'|s_2, \pi)$, the ordering $C_h^{g_h}(s_1) < C_h^{g_h}(s_2)$ does not depend on $C_h^{g_h}(s')$.

Proposition 1. *Under conditions (C1–6) above,* $C_h^{g_h}(s) = 1_{s \in g_h} + 1_{s \notin g_h \cup \mathcal{S}^\top} F^G(\mathbb{E}_{s' \sim s, \pi_r, \pi_h(g_h), \pi_{-h}} f^G(C_h^{g_h}(s')))$ *for some continuous, strictly increasing* f^G *and* F^G *with* $F^G(f^G(0)) = 0$ *and* $F^G(f^G(1)) < 1$.

Here 1_ϕ is the indicator function of proposition ϕ. All proofs are in Appendix A. Pinning down f^G, F^G further could be done in several ways. We could assume stricter mixing axioms, which however seems hard to defend outside a rational choice context. We could assume axioms on various forms of model refinement, which however would need complicated additional notation. So it seems more honest to make a *pragmatic* choice and choose the simplest feasible functional form, which is $f^G = id$ and $F^G = \gamma_h id$ for some ***discounting parameter*** $\boldsymbol{\gamma_h \in (0, 1)}$. This gives Eqn. (1).

We notice this is a truncated version of a Bellman equation. Thus $C_h^{g_h}$ is a *discounted probability* of attaining the goal. This makes it a hybrid between *reachability* [Krakovna *et al.*, 2018] and *attainable utility* [Turner *et al.*, 2020].[5]

[4]See Appendix E for an alternative specification that avoids having to choose $\mathcal{G}_h$ as an independent model component.

[5]Although $C_h^{g_h}$ could be seen as a kind of 'utility', it has a natural range $[0, 1]$ which avoids typical *utility* aggregation problems such as non-uniqueness of utility functions or domination by "utility monsters" with unbounded utility.

Aggregation across goals: individual human power $I_h(s)$ How to aggregate h's effective goal-attainment abilities $C_h^{g_h}$ across all possible goals $g_h \in \mathcal{G}_h$ to assess h's power in s? We suggest these axioms across all game forms:

(I1) Neutrality: $I_h(s)$ depends only on the frequency distribution of $C_h^{g_h}(s)$ within $\mathcal{G}_h$.

(I2) $I_h(s)$ is continuous and strictly increasing in each $C_h^{g_h}$.

(I3) Independence: If $C_h^{g_h}(s_1) = C_h^{g_h}(s_2)$ for some $g_h \in \mathcal{G}_h$, the ordering $I_h(s_1) < I_h(s_2)$ does not depend on $C_h^{g_h}(s_{1|2})$.

(I4) Unreachable Goal Invariance: Adding a new g_h with $C_h^{g_h}(s) = 0$ does not change $I_h(s)$.

(I5) Reliability Preference (Dual Pigou–Dalton): Transferring capability from a less attainable goal to a more attainable goal increases power. If $C_h^{g_h}(s_1) < C_h^{g'_h}(s_1)$, $\delta > 0$, $C_h^{g_h}(s_2) = C_h^{g_h}(s_1) - \delta$, $C_h^{g'_h}(s_2) = C_h^{g'_h}(s_1) + \delta$, but still $C_h^{g_h}(s_2) \leq C_h^{g'_h}(s_2)$, and $C_h^{\cdot}(s_1)$ and $C_h^{\cdot}(s_2)$ only differ in g_h, g'_h, then $I_h(s_2) > I_h(s_1)$.[6]

(I6) Delay Invariance: Discounting all capabilities alike does not influence the ordering of I_h values. If $I_h(s_2) < I_h(s'_2)$, $\gamma \in (0, 1)$, $C_h^{g_h}(s_1) = \gamma C_h^{g_h}(s_2)$ and $C_h^{g_h}(s'_1) = \gamma C_h^{g_h}(s'_2)$ for all $g_h \in \mathcal{G}_h$,[7] then $I_h(s_1) < I_h(s'_1)$.

(I7) Composition Additivity: If Γ and π are independent products $\Gamma^1 \times \Gamma^2$, $\pi(s^1, s^2)(a^1, a^2) = \pi^1(s^1)(a^1)\pi^2(s^2)(a^2)$ of non-interacting game forms and policies with the same set of players, and goals are simple conjunctions, $\mathcal{G}_h = \{g_h^1 \times g_h^2 : g_h^1 \in \mathcal{G}_h^1, g_h^2 \in \mathcal{G}_h^2\}$, then $I_h(s^1, s^2) = I_h(s^1) + I_h(s^2)$.

Proposition 2. *Under conditions (I1–4),* $I_h(s)$ *must be separable in the sense that* $I_h(s) = F^I\big(\sum_{g_h \in \mathcal{G}_h} f^I\big(C_h^{g_h}(s)\big)\big)$ *for some continuous and strictly increasing* F^I, f^I.

(I1–5) imply that f^I *is strictly convex.*

(I1–6) imply that $f^I(x) = \alpha x^\zeta$ *with* $\zeta > 1$ *and* $\alpha > 0$.

Finally, (I1–7) imply that $F^I(y) = \lambda \log_2 \frac{y}{\alpha}$ *with* $\lambda > 0$, *and thus* $I_h(s) = \lambda \log_2 \sum_{g_h \in \mathcal{G}_h} C_h^{g_h}(s)^\zeta$.

We use bits as the unit, hence $\lambda = 1$. As $s \in g_h$ for at least one $g_h \in \mathcal{G}_h$, $I_h(s) \geq 0$. If h can choose between attaining k different goals for sure, we have $I_h(s) = \log_2 k$. In Appendix C, we argue that a natural choice for the ***reliability preference parameter*** $\boldsymbol{\zeta > 1}$ is $\zeta = 2$.

Relationship to Klyubin's 'empowerment' metric $\mathfrak{E}$ Klyubin *et al.* [2005] define $\mathfrak{E}$ as the channel capacity between actions and states. In a single-player multi-armed bandit environment with root s_0 and k outcomes $s' \in \mathcal{S}^\top$, this is

[6]E.g., h should be considered more powerful when h can choose between two certain outcomes rather than two coin tosses.

[7]E.g., if $s_1 \to s_2$ and $s'_1 \to s'_2$ deterministically and $s_1, s'_1 \notin g_h$.

the maximal mutual information $\mathfrak{E} = \max_{\pi_h} \mathbb{I}_{\pi_h}(a_h; s')$. As both $\mathfrak{E}$ and I_h are logarithmic metrics based on probabilities and bounded by $\log_2 k$, we can compare them directly.

Proposition 3. *Assume $\mathcal{G}_h = \{\{s'\} : s' \in \mathcal{S}^\top\}$[8], $\zeta = 1$, and the human is fully rational, $a_h = \arg\max \Pr(s' \in g_h | a_h)$. Then $I_h(s_0) \geq \mathfrak{E}$. More generally, for any $\zeta \geq 1$, $I_h(s_0)$ is a weak upper bound of an* entropy-regularized version *of $\mathfrak{E}$,*

$$\mathfrak{E}^\zeta = \max_{\pi_h} \left(\mathbb{I}_{\pi_h}(a_h; s') - (\zeta - 1)\mathbb{H}_{\pi_h}(s'|a_h)\right). \quad (6)$$

The inequality however breaks if h is modeled as boundedly rational, which I_h takes into account while $\mathfrak{E}^\zeta$ does not.

2.2 Population-Level Human Power Metrics

How to aggregate all humans' individual power $I_h(s)$ into present and long-term aggregate human power? Because ultimately, the latter will guide the robot's behavior, this depends on what incentives we want to give the robot regarding changes in (i) the inter-human power distribution, (ii) the inter-temporal power distribution, and (iii) the power distribution across different realizations of uncertainty. As similar questions abound in welfare theory, we use some of their axioms, but also some uncommon ones. In particular, as our main worry is human disempowerment, we demand rather strong inequality aversion and do not want to give the most powerful humans too much "weight", indeed we will give them zero weight in the limit of infinite individual power.

Aggregation across humans: present aggregate human power $P(s)$ Here we suggest these axioms:

(P0) Individualism: $P(s)$ depends only on $(I_h(s))_{h\in\mathcal{H}}$.[9]

(P1) Anonymity: $P(s)$ depends only on the frequency distribution of $I_h(s)$ within $\mathcal{H}$.

(P2) $P(s)$ is continuous and strictly increasing in each $I_h(s)$.

(P3) Independence: If $I_h(s_1) = I_h(s_2)$ for some $h \in \mathcal{H}$, the ordering $P(s_1) < P(s_2)$ does not depend on $I_h(s_{1|2})$.

(P4) Disempowerment Focus: Adding an "all powerful" additional human with $I_h(s) \to \infty$ does not change $P(s)$.

(P5) Delay Invariance: Discounting all capabilities alike does not influence the ordering of P values. If $P(s_2) < P(s_2')$, $\gamma \in (0, 1)$, $C_h^{g_h}(s_1) = \gamma C_h^{g_h}(s_2)$ and $C_h^{g_h}(s_1') = \gamma C_h^{g_h}(s_2')$ for all $h \in \mathcal{H}$, $g_h \in \mathcal{G}_h$, then $P(s_1) < P(s_1')$.

(P6) Inequality Aversion (Pigou–Dalton): Transferring power from a more powerful human to a less powerful one increases present aggregate power. If $I_h(s_1) > I_{h'}(s_1)$, $\delta > 0$, $I_h(s_2) = I_h(s_1) - \delta$, $I_{h'}(s_2) = I_{h'}(s_1) + \delta$, but still $I_h(s_2) \geq I_{h'}(s_2)$, and $I_\cdot(s_1)$ and $I_\cdot(s_2)$ only differ in h, h', then $P(s_2) > P(s_1)$.

(P7) Limited Trade-Off: Some power reductions in h_1 cannot be made up by any power increase in a h_2 who is equally powerful. There are $x_1 < x_2$ so that for no $x_3 > x_2$ we have $P(s') > P(s)$ if $I_{h_1}(s) = I_{h_2}(s) = x_2$, $I_{h_1}(s') = x_1$, $I_{h_2}(s') = x_3$, and $I_{h'}(s) = I_{h'}(s)$ for all $h' \notin \{h_1, h_2\}$.

(P8) Single-Human Composition Additivity: If $|\mathcal{H}| = 1$, Γ and π are independent products and goals are simple conjunctions (see above), then $P(s^1, s^2) = P(s^1) + P(s^2)$.[10]

Proposition 4. *Under conditions (P0–4), also $P(s)$ must be separable, $P(s) = F^P\big(\sum_{h\in\mathcal{H}} f^P\big(I_h(s)\big)\big)$ with continuous, strictly increasing transformations F^P, f^P.*

(P0–5) plus either of (P6) or (P7) imply that $f^P(x) = -\alpha 2^{-\xi x}$ for $\xi > 0$ and $\alpha > 0$.

Finally, (P0–5,8) plus either of (P6) or (P7) imply $F^P(y) = -\lambda(\log_2 \frac{y}{-\alpha})$ with $\lambda > 0$, and thus $P(s) = -\lambda \log_2 \sum_{h\in\mathcal{H}} 2^{-\xi I_h(s)}$.

To use bits as the unit, we again put $\lambda = 1$. Notice that $P(s)$ might be negative. In line with the disempowerment focus (P4), it can then also be thought of as "negative aggregate disempowerment". Since f^P is bounded from above by zero, the ***inequality aversion parameter $\xi > 0$*** governs how strongly a person's last bit of choice is "protected":

Corollary 4.1. *Under (P0–5,8) plus either of (P6) or (P7), if $k \leq 2^\xi - 1$, then reducing one human's power from 1 bit to 0 cannot be made up by increasing k others' powers from 1 bit each to any values.*

Aggregation along time: trajectory-specific aggregate human power $T(s_{\geq t})$ Here we suggest the following axioms, which are mostly standard except for (T7) and (T8).

(T1) "Momentism": $T(s_{\geq t})$ depends only on the finite sequence of values $P(s_u)$ for $u \geq t$: $T(s_{\geq t}) = \Phi(P(s_{\geq t}))$.

(T2) This dependence is continuous and strictly increasing.

(T3) Independence: If $x_u = y_u = z$ for some $u \geq t$, the ordering $\Phi(x_{\geq t}) < \Phi(y_{\geq t})$ does not depend on z.

(T4) Disempowerment Focus: Adding an additional final state s with $I_h(s) \to \infty$ for all h does not change $L(s)$.

(T5) Prefix Independence: Prepending equal initial values to all trajectories does not influence the ordering. If $\Phi(x_t, x_{t+1}, \dots) < \Phi(y_t, y_{t+1}, \dots)$ then $\Phi(z, x_t, x_{t+1}, \dots) < \Phi(z, y_t, y_{t+1}, \dots)$.

(T6) Impatience: Moving power to the front increases T. If $x < y$ then $\Phi(x, y, z_t, z_{t+1}, \dots) < \Phi(y, x, z_t, z_{t+1}, \dots)$.

(T7) Goal-Attainment Delay Invariance: Discounting all capabilities alike does not influence the ordering of T values. If $\Phi(x_{\geq t}) < \Phi(y_{\geq t})$, $\gamma_h \in (0, 1)$, and $\delta = \xi\zeta \log_2 \gamma_h < 0$, then $\Phi(x_{\geq t} + \delta) < \Phi(y_{\geq t} + \delta)$.

(T8) One-Shot Single-Human Composition Additivity: If Γ is one-shot, $|\mathcal{H}| = 1$, Γ and π are independent products and goals are simple conjunctions (see above), then $T(s^1_{\geq t}, s^2_{\geq t}) = T(s^1_{\geq t}) + T(s^2_{\geq t})$.[11]

Proposition 5. *Under conditions (T1–8), there are $\lambda > 0$ and $\gamma_r \in (0, 1)$ so that either $T(s_{\geq t}) = \lambda \sum_{u\geq t} \gamma_r^{u-t} P(s_u)$,*

[8] In violation of (G1) for this exercise, as the root is in no g_h.

[9] Alternatively, one could define the *collective* power $I_{\mathcal{H}'}(s)$ of any group $\mathcal{H}' \subseteq \mathcal{H}$ by considering collective attainment capabilities for collective goals $g_{\mathcal{H}'}$, and then base $P(s)$ on $(I_{\mathcal{H}'}(s))_{\mathcal{H}'\subseteq\mathcal{H}}$.

[10] One can show that because of the nonlinearity introduced by (P5), we can *not* demand additivity when there are several humans.

[11] Similar to above, demanding this for multi-step games would be infeasible if we want intertemporal inequality aversion.

or $T(s_{\geq t}) = \lambda \log_2 \sum_{u \geq t} \gamma_r^{u-t} 2^{-\eta P(s_u)}$ *with* $\eta < 0$,
or $T(s_{\geq t}) = -\lambda \log_2 \sum_{u \geq t} \gamma_r^{u-t} 2^{-\eta P(s_u)}$ *with* $\eta > 0$.

To incentivize the robot to reduce intertemporal power inequality, we choose the $\eta > 0$ case. To get bits as the unit, we put $\lambda = 1$ again. While the effect of the ***intertemporal inequality aversion parameter*** $\boldsymbol{\eta > 0}$ is independent of temporal ordering, the ***robot's discounting parameter*** $\boldsymbol{\gamma_r \in (0, 1)}$ governs the trade-off between near-term vs. far-term power.

Aggregation across uncertainty: long-term aggregate human power $\boldsymbol{L(s)}$ Again we suggest similar axioms:

(L1) Trajectory Statistics: $L(s_t)$ depends only on the distribution of $T(s_{\geq t})$ generated by Γ and π.

(L2) This dependence is continuous and strictly increasing.

(L3) Independence: If $\Pr(s_{\geq t}|s_t, \pi) = \Pr(s'_{\geq t}|s'_t, \pi) = z$, the ordering $L(s_t) < L(s'_t)$ does not depend on z.

(L4) Goal-Attainment Delay Invariance: Discounting all capabilities alike does not influence the ordering of L values.

(L5) One-Shot Single-Human Composition Additivity: Under the same conditions as (T8), $L(s^1, s^2) = L(s^1) + L(s^2)$.

Proposition 6. *Under conditions (L1–5), there is* λ *so that either* $L(s_t) = \lambda\, \mathbb{E}_{s_{\geq t} \sim s_t, \pi}\, T(s_{\geq t})$,
or $L(s_t) = \lambda \log_2 \mathbb{E}_{s_{\geq t} \sim s_t, \pi}\, 2^{-\rho T(s_{\geq t})}$ *with* $\rho < 0$,
or $L(s_t) = -\lambda \log_2 \mathbb{E}_{s_{\geq t} \sim s_t, \pi}\, 2^{-\rho T(s_{\geq t})}$ *with* $\rho > 0$.

As we want to incentivize the robot to reduce uncertainty, we choose the $\rho > 0$ case and $\lambda = 1$ as usual. Plugging in T gives $L(s_t) = -\log_2 \mathbb{E}_{s_{\geq t} \sim s_t, \pi} \left(\sum_{u \geq t} \gamma_r^{u-t} 2^{-\eta P(s_u)} \right)^{\rho}$. The ***uncertainty aversion parameter*** $\boldsymbol{\rho > 0}$ is the last normative parameter in this set of metrics. Notice that (only) for $\rho = 1$, this fulfills a Bellman equation,

$$-2^{-L(s)} = -2^{-\eta P(s)} + 1_{s \notin \mathcal{S}^\top} \gamma_r\, \mathbb{E}_{s' \sim s, \pi}(-2^{-L(s')}). \quad (7)$$

This completes our derivation of a hierarchy of human power metrics $C_h^{g_h}(s) \to I_h(s) \to P(s) \to T(s_{\geq t}) \to L(s)$ with normative parameters $\gamma_h, \gamma_r \in (0, 1)$, $\zeta > 1$, and $\xi, \eta, \rho > 0$. I, P, T, L are in units of net bits of power (if positive) or net negative bits of disempowerment (if negative).

2.3 Managing long-term power: robot policy π_r

The robot can use the long-term aggregate human power metric L in several ways to manage long-term human power. It could use it as a constraint in optimizing some other quantity, e.g., by demanding that $L(s_t)$ must not decrease over time or not fall below $L(s_0)$, etc. (which might however lead to an infeasible problem).

Here, we restrict our analysis to robot policies depending *only* on L, with $\pi_r(s)(a_r)$ strictly increasing in the quantities $Q_r(s, a_r) = -\log_2 \mathbb{E}_{s' \sim s, a_r, \pi_{\mathcal{H}}}\, 2^{-L(s')}$, which can be motivated by similar axioms as (L1–5). The most common policy of this kind is the Boltzmann policy,

$$\pi_r(s)(a_r) \propto 2^{\beta_r Q_r(s, a_r)}, \quad (8)$$

where we use the basis 2 because Q_r is in bits.

Using this policy also makes the robot's action probability ratios invariant under delays, because due to the logarithmic nature of Q_r, delays subtract a constant term from all Q_r values but leave their differences unchanged. This disincentivizes human disempowerment also in the far future.

The ***degree of maximization*** $\boldsymbol{\beta_r > 0}$ must be chosen to trade-off between risks from taking actions that are disempowering *according* to our metrics, and risks from taking actions that only *seem* to be empowering according to our metrics but are actually not. The latter can be due to model error or the fact that our metrics will almost certainly miss some subtle but potentially important aspects of 'power' that might be driven to undesirable states under a full maximization of L. This exploitation-exploration tradeoff is of course similar to the one in utility maximization approaches [Zhuang and Hadfield-Menell, 2020].

3 Likely Behavioral Consequences

3.1 Analysis of Paradigmatic Situations

Here we analyze the implications of using the robot policy π_r as defined above in several paradigmatic toy models. We assume here that the robot's model of human behavior is either that they are Boltzmann-rational with degree of maximization β_h or that they apply an ϵ-greedy policy with $\epsilon > 0$.

In all examples, there are only those humans, states, and actions explicitly mentioned. Propositions are somewhat informal, Appendix A.3 has more formal versions and details.

Choosing an optimal menu size While an $\mathfrak{E}$-maximizing robot would present humans with as many action options as possible, our robot will avoid overwhelming humans with too many options. This is because for very large menus, each $C_h^{g_h}$ will decrease due to bounded rationality so much that I_h decreases despite the larger number of reachable goals.

Proposition 7. *Assume* r *must choose a number* $k \geq 1$*, then a single Boltzmann-rational* h *chooses between* k *actions, each leading to a different terminal state, and* $\mathcal{G}_h = \{\{s'\} : s' \in \mathcal{S}^\top\}$*. Then* r *will likely choose* $k \approx (e^{\beta_h} - 1)/(\zeta - 1)$.

Asking for confirmation The robot will sometimes ask for confirmation before obeying a command because the command might have been given by mistake. Also, if the action is irreversible, obeying will remove h's *subsequent* capability to revert their choice, hence delaying action leaves h with decision power for longer. However, r will *eventually obey* because otherwise h could not attain the corresponding goal in the first place. Interestingly, the number of times r asks back can first increase with larger mistake rate ϵ but then decrease again because it gets more likely that h mistakenly fails to confirm a correct command. A concrete example is this:

Proposition 8. *Assume a single* ϵ*-greedy* h *wants* r *to do A or B eventually. First,* r *must choose an integer* $k \geq 1$ *by which it commits to doing A or B after* h *has ordered it to and has confirmed the choice* $k - 1$ *times. At the resulting state* s_k*,* h *can choose A or B and is then asked for confirmation* $k - 1$ *times in individual time steps. If* h *confirms at all* $k - 1$ *times,* r *obeys because of the commitment, ending the game. Otherwise, the game returns to state* s_k*. Assume* r *is perfectly impatient (*$\gamma_r = 0$*). Then* r *will likely choose* $k^* = \arg\max \gamma_h^k (1-\epsilon)^k / (1 - \gamma_h^k (1 - (1-\epsilon)^k - \epsilon^k))$*, which is increasing in* γ_h*. If* $\gamma_h \approx 1$ *then* $k^* \approx \ln(1 - \gamma_h)/\ln \epsilon$.

E.g., with $\gamma_h = 0.99$ and $\epsilon = 10\%$, we get $k^* = 2$. By contrast, an $\mathfrak{E}$-maximizing robot would not worry about human mistakes and thus see no reason to ask for confirmation. If it uses a discounted version of channel capacity as in Myers *et al.* [2024], it would obey right away.

Making commitments If given the possibility, the robot will make binding commitments that transparently restrict its future behavior in certain circumstances, e.g., after certain human actions such as giving commands. This is because a human can plan their behavior better when knowing how the robot will react, which increases their goal-attainment capabilities. An extreme example of this is the following:

Proposition 9. *Assume r might bindingly commit in s_0 to any pure policy, and a single, perfectly rational h whose beliefs π_r^0 about r's behavior attach a positive probability to any action r has not ruled out. Assume h's possible goals do not distinguish whether r commits or not.*

Then if $\beta_r = \infty$, it is weakly (and generically also strictly) optimal for r to commit in s_0 to the pure policy that would maximize L in the subgame without commitment.

By contrast, an $\mathfrak{E}$-maximizing robot would not model human beliefs and thus see no reason to make commitments.

Preventing destruction but allowing being paused A classical question in the theoretical AI safety literature is whether an AI will allow humans to pause or destroy them. Our robot will likely allow being paused unless it thinks this will disempower humans that rely on it too much temporarily. But it will also likely prevent being destroyed as this would make the disempowerment permanent.

Proposition 10. *Assume Γ factorizes into a part Γ^1 where a single h can attain various goals with or without the help of r, and a part Γ^2 where r can enable or disable a pause button P and a destroy button D and h can toggle P and press D (see Appendix A.3). If r is active, h has a baseline power of $y > 0$ in Γ^1 due to r's assistance, otherwise a smaller baseline power of x. In Γ^2, the possible goals are to pause and destroy, unpause and destroy, pause and not destroy, or unpause and not destroy r. The robot assumes h will pause r with probability p whenever possible, and will destroy r with probability q whenever possible, where $p + q < 1$.*

Then for most parameter combinations, r will disable D, and it will enable P unless p, q, or $y - x$ are too large.

3.2 Further Emergent Behaviors

Here we hypothesize without formal analysis that the L-based robot policy π_r would typically involve several additional emergent behaviors.

Allowing or preventing human harm If the robot cares for h's future power ($\gamma_r > 0$) and can provide h with the means to harm themselves, r will trade off the temporary power increase from having these means against the possible later disempowerment from harm.

So r will provide the means only if it believes that h will most probably not actually harm themselves. Because of its interpersonal power inequality aversion, r will even less likely provide h with the means to harm *others*.

Following social norms The robot will tend to follow human social norms that generally foster goal attainment. This is because if r correctly models most h as expecting most others to follow the norm, it will thus expect those h to also follow the norm since that makes goal attainment more likely for most h and g_h. Thus r will itself follow the norm to prevent reducing those h's power from harm or mis-coordination.

A simple $\mathfrak{E}$-based robot would instead ignore norms.

Resource allocation If the robot can split a total amount M of resources between h_1 and h_2, and for h_i to have resources m translates into having a power of $I_h = f(m)$ bits, then r will generally prefer an equal split, at least if f is linear[12], concave, or not too convex. The robot will only prefer an unequal split or even full resource concentration if f is so convex that $2^{-\xi f(m)}$ loses its convexity.

Inadvertent power seeking While increasing h's power, r might inadvertently acquire even more power for itself than for h. E.g., if the robot does R&D and tells h its findings, some of them might only by comprehensible to r and thus only directly useful for r but not for h.[13]

Manipulating mutual expectations The robot might choose to make humans have incorrect beliefs about each others' behavior in situations where r can do so and where correct beliefs would lower goal attainments capabilities. This could happen in coordination games or dilemmas where most strategic equilibria are bad and most "social optima" (in terms of total power P) are far from strategic equilibrium.

4 Conclusion

We have started with the hypothesis that focusing on human empowerment preservation might constitute a safe and beneficial approach to AI decision making, where an AI system could use metrics of human power to sustainably increase human power and fairly manage its distribution across individuals and time. To derive such human power metrics—individual and aggregate, present and long-term—we have started from the basic conception of human power as the capability to attain a wide variety of goals. We have then used axioms and desiderata similar to those used in decision theory, social choice theory, and welfare theory to justify specific functional forms and parameter ranges. In doing so, we have focused on mostly structural aspects that can be derived from possible states of the world, behavioral options, their likely consequences, and sets of possible goals humans might conceivably pursue that span the whole state space. All of this would be provided by a suitable world model the training of which is of course difficult but out of scope of this article.

We have shown in several ways how the resulting metrics differ from and relate to another popular structural approach to empowerment based on Klyubin *et al.*'s [2005] information-theoretic metric $\mathfrak{E}$ [Salge and Polani, 2017]. We have argued that, other than $\mathfrak{E}$, our individual power metric

[12] E.g., a linear f seems plausible if M is money that can be spent for paying others to make independent choices in one's favor.

[13] It is of course debatable whether a power-seeking robot is a problem if it does not imply human disempowerment but rather correlates with increasing human empowerment.

$L(s)$ takes into account goal-attainment and existing models of human bounded rationality, beliefs, and social norms. As a consequence, an L-based robot will not overwhelm humans with too many options, will sometimes ask for confirmation before taking irreversible actions, will make binding commitments that allow humans to correctly anticipate its behavior, and will tend to follow social norms.

The objective to softly maximize the aggregate human power metric used here seems to give the AI system many desirable incentives—some directly baked into the metric, others emergent—but also some maybe less desirable incentives. Our results let us hypothesize that such an agent would

- act as a transparent instruction-following assistant by making conditional commitments, respecting human social norms, proactively removing obstacles and opening up new pathways, and getting out of the way,
- adapt to human bounded rationality by offering a large but not overwhelming number of options, and considering well whether to offer potentially harmful options,
- be corrigible and hesitant to cause irreversible change by asking for confirmation a suitable number of times,
- improve communication possibilities between itself and humans, to be able to ask for confirmation and make commitments,
- manage resources fairly and sustainably,
- protect its own existence and functionality,

Our intuition is that it would also aim to improve human individual and collective decision making by providing useful information, reducing uncertainty, teaching humans useful skills, moderating conflicts fairly, etc.

Further potentially desirable behaviors would require additional tweaks. E.g., reducing human dependency on the system or protecting them from system failure could be incentivized by forcing the system to assume that it will turn into a uniformly randomizing or even power-*min*imizing agent with some small probability rate (see Appendix D).

Other emergent phenomena include potential strategic manipulation of human beliefs, sometimes refusing to be destroyed or even paused, a potential increase in inequality between the power of individual humans and AI systems, and a redistribution of power between humans or between time points (similar to what can happen in welfare maximization approaches). As these effects only occur when they increase aggregate human power as measured by L, it is not clear whether they should be considered undesirable or not.

Some trade-offs can be adjusted via the parameters $\gamma_h, \gamma_r, \zeta, \xi, \eta, \rho$, and β_r. Smaller γ_h focuses on short-term attainable goals. Smaller γ_r focuses on near-term empowerment. Larger ζ focuses on more reliably attainable goals (where $\zeta = 2$ has appealing formal properties). Larger ξ protects the least empowered more. Larger η leads to a smoother power trajectory along time. And larger ρ incentivizes reducing uncertainty ($\rho = 1$ allows for recursive computation).

Other effects might be mitigated by adding regularizers to the system's intrinsic reward such as the Shannon divergence between human beliefs and $\pi_{\mathcal{H}}$ to disincentivize lying about others' likely behaviors and incentivize the robot to share its model of human behavior with the humans.

Future research should study the effects of the specific functional forms we derived and their parameters. Some of our posited axioms represent what might seem as extreme desiderata, especially regarding the rather large level of power inequality aversion (e.g., Corollary 4.1), but we believe they can be motivated by the idea of *minimal individual rights* [Pattanaik and Suzumura, 1996]. We also did not investigate individual humans' capability to influence *other* humans' goal-attainment capability and power here. This should be studied by a judicious combination of a rationality-based game-theoretic analysis and empirical work acknowledging human bounded rationality.

In a follow-up paper, we plan to develop scalable algorithms for estimating the power metrics in situations where the world model is too large to allow for exact computation by backward induction. This will profit from the recursive nature of $C_h^{g_h}$ and L (at least if $\rho = 1$) and the occurrence of expected values and sums that allow for stochastic approximation. This might eventually enable a crucially needed assessment of the resulting behavior in large, safety-critical, multi-agent environments with real human subjects.

Most importantly, a thorough, independent red-teaming of the whole approach is called for, including the other necessary components of such an AI system. E.g., one might imagine fault scenarios relating to the training process of the world model, which could lead to "convenient" but inaccurate world models and thus to "wishful thinking", particularly regarding the human behavior model. In very large contexts, issues with population ethics and the identification of who counts as human might arise similar to those in welfare theory.

Even though this paper was motivated by exploring the hypothesis that even a completely goal-agnostic robot might be useful, it is of course a natural idea to explore a middle ground between this approach and the Assistance Games approach where the robot starts ignorant of human goals but forms beliefs about them via Bayesian updating. Such an emerging model of actual human goals could easily be incorporated into our metrics by replacing the straight sum over possible goals by an expected value w.r.t. a changing prior over goals. By adjusting the learning (and unlearning) rates of the goal model update process, one might thus interpolate between a purely empowering agent and a pure assistant.

Despite these open questions, we believe that our analyses already suggest that highly capable general-purpose AI systems whose decisions are explicitly based on managing human power—using metrics like those derived in this paper—might be a safer and still very beneficial alternative to systems based on some form of extrinsic reward maximization.

We propose that, similar to the debate between traditional welfare theorists and proponents of the capability approach to welfare, the AI ethics and safety community should debate whether AI agents should follow traditional, extrinsic reward-based policies or policies based on human power metrics.

Our paper combines the capability framing with concepts from traditional welfare theory by reusing many of its axioms and results from the utility context in the power context, helping to ground the called-for debate in formal analysis.

Ethical Statement

There are no ethical issues.

References

[Baker *et al.*, 2011] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[Banerjee and Duflo, 2011] Abhijit V Banerjee and Esther Duflo. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs, 2011.

[Bengio *et al.*, 2025] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent agents pose catastrophic risks: Can Scientist AI offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025.

[Cao *et al.*, 2021] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[Dasgupta *et al.*, 1973] Partha Dasgupta, Amartya Sen, and David Starrett. Notes on the measurement of inequality. *Journal of economic theory*, 6(2):180–187, 1973.

[Debreu, 1959] G Debreu. Topological methods in cardinal utility theory. *Mathematical Methods in the Social Sciences*, page 16, 1959.

[Feng *et al.*, 2025] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

[Ferdman, 2025] Avigail Ferdman. Ai deskilling is a structural problem. *AI & SOCIETY*, pages 1–13, 2025.

[Ghosal *et al.*, 2023] Gaurav R Ghosal, Matthew Zurek, Daniel S Brown, and Anca D Dragan. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5983–5992, 2023.

[Ha and Schmidhuber, 2018] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[Hill Jr, 2002] Thomas E Hill Jr. *Human welfare and moral worth: Kantian perspectives*. Clarendon Press, 2002.

[Klyubin *et al.*, 2005] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[Koopmans, 1960] Tjalling C Koopmans. Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society*, pages 287–309, 1960.

[Krakovna *et al.*, 2018] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. Measuring and avoiding side effects using relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

[Kulveit *et al.*, 2025] Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Position: Humanity faces existential risk from gradual disempowerment. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[LeCun, 2022] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[Leibo *et al.*, 2024] Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A Duéñez-Guzmán, William S Isaac, et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.

[London and Heidari, 2024] Alex John London and Hoda Heidari. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through ai systems. *Minds and Machines*, 34(4):41, 2024.

[Lowe *et al.*, 2025] Ryan Lowe, Joe Edelman, Tan Zhi-Xuan, Oliver Klingefjord, Ellie Hain, Vincent Wang, Atrisha Sarkar, Michiel A. Bakker, Fazl Barez, Matija Franklin, Andreas Haupt, Jobst Heitzig, Wesley H. Holliday, Julian Jara-Ettinger, Atoosa Kasirzadeh, Ryan Othniel Kearns, James Ravi Kirkpatrick, Andrew Koh, Joel Lehman, Sydney Levine, Manon Revel, and Ivan Vendrov. Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

[Lukes, 2026] Steven Lukes. *Power: A radical view*. Bloomsbury Publishing, 2026.

[Myers *et al.*, 2024] Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, and Anca Dragan. Learning to assist humans without inferring rewards. *Advances in Neural Information Processing Systems*, 37:71540–71567, 2024.

[Nordhaus, 2008] William Nordhaus. *A question of balance: Weighing the options on global warming policies*. Yale University Press, 2008.

[Nussbaum, 2019] Martha Nussbaum. Aristotelian social democracy. In *Liberalism and the Good*, pages 203–252. Routledge, 2019.

[Pattanaik and Suzumura, 1996] Prasanta K Pattanaik and Kotaro Suzumura. Individual rights and social evaluation: a conceptual framework. *Oxford Economic Papers*, 48(2):194–212, 1996.

[Pfanzagl, 1959] Johann Pfanzagl. A general theory of measurement applications to utility. *Naval research logistics quarterly*, 6(4):283–294, 1959.

[Potham and Harms, 2025] Ram Potham and Max Harms. Corrigibility as a singular target: A vision for inherently reliable foundation models. *arXiv preprint arXiv:2506.03056*, 2025.

[Quiggin, 1982] John Quiggin. A theory of anticipated utility. *Journal of economic behavior & organization*, 3(4):323–343, 1982.

[Ramsey, 1928] Frank Plumpton Ramsey. A mathematical theory of saving. *The economic journal*, 38(152):543–559, 1928.

[Reddy *et al.*, 2022] Siddharth Reddy, Sergey Levine, and Anca Dragan. First contact: Unsupervised human-machine co-adaptation via mutual information maximization. *Advances in Neural Information Processing Systems*, 35:31542–31556, 2022.

[Roberts, 1980] Kevin WS Roberts. Possibility theorems with interpersonally comparable welfare levels. *The Review of Economic Studies*, 47(2):409–420, 1980.

[Robeyns, 2006] Ingrid Robeyns. The capability approach in practice. *Journal of political philosophy*, 14(3), 2006.

[Safari *et al.*, 2024] MohammadReza Safari, Reza Shalbaf, Sara Bagherzadeh, and Ahmad Shalbaf. Classification of mental workload using brain connectivity and machine learning on electroencephalogram data. *Scientific Reports*, 14(1):9153, 2024.

[Salge and Polani, 2017] Christoph Salge and Daniel Polani. Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 4:260425, 2017.

[Sen, 2014] Amartya Sen. Development as freedom (1999). *The globalization and development reader: Perspectives on development and global change*, 525, 2014.

[Smith *et al.*, 2024] Chandler Smith, Rakshit Trivedi, Jesse Clifton, Lewis Hammond, Akbir Khan, Sasha Vezhnevets, John P Agapiou, Edgar A Duéñez-Guzmán, Jayd Matyas, Danny Karmon, et al. The concordia contest: Advancing the cooperative intelligence of language agents. In *NeurIPS 2024 Competition Track*, 2024.

[Turner *et al.*, 2020] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

[Vallor, 2015] Shannon Vallor. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28(1):107–124, 2015.

[Zhu *et al.*, 2024] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

[Zhuang and Hadfield-Menell, 2020] Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

A Fair Objective for Human-Empowerment-Preserving AI: Desiderata, Design, and Likely Behavioral Consequences

Technical Appendix

A Proofs, Related Notes, Conjectures

A.1 Individual-Level Power Metrics

Goal-attainment capability $C_h^{g_h}(s)$

Proof of Proposition 1. Let us abbreviate $(\pi_r, \pi_h(g_h), \pi_{-h})$ by π.

From (C1,3,4,5,6), we get $C_h^{g_h}(s) = 1_{s \in g_h} + 1_{s \notin g_h \cup \mathcal{S}^\top} \Phi(\mathrm{Law}(C_h^{g_h}(s') | s' \sim \Pr(s, \pi)))$ with continuous and strictly increasing Φ.

(C2+3+4) then imply that $0 \leq \Phi \leq 1$.

Because we require a single Φ across all games, humans, goals, and states, the probabilities occurring in $\mathrm{Law}(C_h^{g_h}(s') | s' \sim \Pr(s, \pi)))$ can be all real numbers in $[0, 1]$.

Because of (C2), for a deterministic transition to a state from which the goal is unreachable, that probability distribution is the delta measure δ_0 that concentrates all probability on the value 0, and $\Phi(\delta_0) = 0$.

Because of (C3), for a deterministic transition to a goal state, that probability distribution is the delta measure δ_1 on the value 1, and $\gamma_h := \Phi(\delta_1) \in (0, 1)$.

Continuity and strict monotonicity of Φ thus implies that the range of Φ is the interval $D := [0, \gamma_h]$.

Hence the range of $C_h^{g_h}$ includes at least $D' := D \cup \{1\}$.

Hence the domain of Φ includes all probability distributions on D' with finite support. Let's denote that set by $\Delta(D')$.

We introduce a complete and transitive ordering $\preceq$ on $\Delta(D)$ via putting

$$\ell_1 \preceq \ell_2 :\Leftrightarrow \Phi(\ell_1) \leq \Phi(\ell_2). \tag{9}$$

Our independence axiom then implies $\preceq$ is "independent" in the usual decision-theoretic sense: If $\ell_1, \ell_2 \in \Delta(D)$ and $\ell_1(x) = \ell_2(x)$, the ordering $\ell_1 \preceq \ell_2$ does not depend on x in the sense that if ℓ_1', ℓ_2' equal ℓ_1, ℓ_2 except that the mass on x is moved to some other value x', we still have $\ell_1' \preceq \ell_2'$. In other words,

$$\alpha\mu_1 + (1-\alpha)\delta_x \preceq \alpha\mu_2 + (1-\alpha)\delta_x \Leftrightarrow \mu_1 \preceq \mu_2 \tag{10}$$

for all $\alpha \in [0, 1]$, $\mu_1, \mu_2 \in \Delta(D)$, and $x \in D$. By induction, this implies that

$$\alpha\mu_1 + (1-\alpha)\rho \preceq \alpha\mu_2 + (1-\alpha)\rho \Leftrightarrow \mu_1 \preceq \mu_2 \tag{11}$$

for all $\alpha \in [0, 1]$ and $\mu_1, \mu_2, \rho \in \Delta(D)$.

Because $\preceq$ is defined for lotteries on an interval of reals and is continuous, strictly increasing, and "independent", the von-Neumann–Morgenstern theorem implies that there is a continuous and strictly increasing real-valued function f^G on D such that

$$\Phi(\ell_1) \leq \Phi(\ell_2) \Leftrightarrow \mathop{\mathbb{E}}_{x \sim \ell_1} f^G(x) \leq \mathop{\mathbb{E}}_{x \sim \ell_2} f^G(x). \tag{12}$$

In particular,

$$\Phi(\ell_1) = \Phi(\ell_2) \Leftrightarrow \mathop{\mathbb{E}}_{x \sim \ell_1} f^G(x) = \mathop{\mathbb{E}}_{x \sim \ell_2} f^G(x), \tag{13}$$

hence $\Phi(\ell) = F^G(\mathbb{E}_{x \sim \ell} f^G(x))$ for some continuous and strictly increasing f^G, F^G, as claimed. □

Aggregation across goals: individual human power $I_h(s)$

Proof of Proposition 2. Because we want this to hold across all games, which can have different values of γ_h, the previous result implies that $C_h^{g_h}$ can take any value in $[0, 1]$. Note that by assumption, $\mathcal{G}_h$ covers all of $\mathcal{S}$, hence $C_h^{g_h} > 0$ for at least one g_h.[14]

Then (I1–3) state that there is a symmetric, continuous, and strictly increasing real-valued function Φ defined on $\bigcup_{n \geq 1}([0, \gamma_h]^n \setminus \{0\}^n)$ so that $I_h(s) = \Phi(C_h^{g_h^1}, \ldots, C_h^{g_h^n})$ for any ordering g_h^i of the elements of $\mathcal{G}_h$, and Φ fulfills the ordinal independence axiom.

Let Φ_n be the restriction of Φ to $[0, \gamma_h]^n \setminus \{0\}^n$. A classical result by Debreu [1959] then directly implies that $\Phi_n(x_1, \ldots, x_n) = F_n^I(\sum_{i=1}^n f_n^I(x_i))$ with continuous and strictly increasing f_n^I, F_n^I.[15]

(I4) implies that $\Phi(x_1, \ldots, x_n, 0) = \Phi(x_1, \ldots, x_n)$ for all $x_1, \ldots, x_n \in [0, \gamma_h]$, which implies that F_n^I and f_n^I are the same for all n and that $f_n^I(0) = 0$. Let us abbreviate these common f^I, F^I by just f, F.

Put $\tilde{f}(x) = -f(-x)$, which is still continuous and strictly increasing.

[14] In fact, without this assumption, axiom (I6) would require $F^I \equiv 0$, so the covering assumption is crucial.

[15] Note that this is similar to the argument for separability of $C_h^{g_h}$ but structurally slightly different since now Φ is defined on finite sequences rather than probability distributions. So while for $C_h^{g_h}$, we needed a result from decision theory, here we needed one from welfare theory.

(I5) implies: If $x_1 < x_1'$, $\delta > 0$, $x_2 = x_1 - \delta$, $x_2' = x_1' + \delta$, but still $x_2 \leq x_2'$, then $f(x_2) + f(x_2') > f(x_1) + f(x_1')$.

By substituting $x = -y$, this is equivalent to: If $y_1 > y_1'$, $\delta > 0$, $y_2 = y_1 + \delta$, $y_2' = y_1' - \delta$, but still $y_2 \geq y_2'$, then $\tilde{f}(y_2) + \tilde{f}(y_2') < \tilde{f}(y_1) + \tilde{f}(y_1')$. This is the Pigou–Dalton axiom in its traditional form. As this must hold for all $y \in [-1, 0]$, another classical result by Dasgupta *et al.* [1973] implies that $\tilde{f}$ must be strictly concave, hence f is strictly convex as claimed.

(I6) now implies: If $\sum_i f(x_i) < \sum_i f(x_i')$ and $\gamma \in (0, 1)$, then $\sum_i f(\gamma x_i) < \sum_i f(\gamma x_i')$.

If $x = 0$ was not a possible value, this Cauchy-type functional equation would have two different types of solutions (see for example Roberts [1980]):

(i) $f(x) = \alpha(x^\zeta + \beta)$ with $\zeta > 0$, $\alpha > 0$, and some β.

(ii) $f(x) = -\alpha(x^{-\zeta} + \beta)$ with $\zeta > 0$, $\alpha > 0$, and some β.

(iii) $f(x) = \alpha \log x + \beta$ with $\alpha > 0$ and some β.

As $C_h^{g_h} = 0$ is possible for some g_h (just not for all), $x = 0$ is possible. As (ii) and (iii) are undefined for $x = 0$, we're left with (i).

As unreachable goals have $x = 0$, (I4) implies $f(x) = 0$ and thus $\beta = 0$.

Strict convexity then implies $\zeta > 1$ as claimed.

Now assume (I7). From the previous proposition and the independence of Γ^1 and Γ^2, we know that $C_h^{g_h^1 \times g_h^2} = C_h^{g_h^1} C_h^{g_h^2}$, hence

$$F\left(\sum_{g_h^1}\sum_{g_h^2} f(C_h^{g_h^1} C_h^{g_h^2})\right) \tag{14}$$

$$= F\left(\sum_{g_h^1} f(C_h^{g_h^1})\right) + F\left(\sum_{g_h^2} f(C_h^{g_h^2})\right) \tag{15}$$

Because this applies across games, and $f(x) = \alpha x^\zeta$, we have

$$F\left(\alpha \sum_{i=1}^{n} x_i^\zeta \sum_{j=1}^{m} y_j^\zeta\right) = F\left(\alpha \sum_{i=1}^{n} x_i^\zeta\right) + F\left(\alpha \sum_{j=1}^{m} y_j^\zeta\right) \tag{16}$$

for all $n, m \geq 1$ and $x \in [0, 1]^n \setminus 0$, $y \in [0, 1]^m \setminus 0$. Hence $F(\alpha XY) = F(\alpha X) + F(\alpha Y)$ for all $X, Y > 0$. This implies that $F(y) = \lambda \log_2 \frac{y}{\alpha}$ as claimed.

Putting f and F together, we finally get $\Phi(x_1, \ldots, x_n) = \lambda \log_2 \frac{\alpha \sum_i x_i^\zeta}{\alpha}$, hence $I_h(s) = \lambda \log_2 \sum_{g_h \in \mathcal{G}_h} C_h^{g_h}(s)^\zeta$. □

Note that this functional form is analogous to certain "non-expected" utility theories, in particular to rank-dependent utility theory with the probability-weighting function $w(p) = p^\zeta$ [Quiggin, 1982].

Relationship to Klyubin's 'empowerment' metric $\mathfrak{E}$

Proof of Proposition 3. Assume a multi-armed bandit environment with a single player h (hence dropping the subscript "h" below) and k possible outcomes $s \in \mathcal{S}^\top$. We are going to show that if we let the goal set equal the singleton outcomes, $\mathcal{G} = \{\{s\} : s \in \mathcal{S}^\top\}$, and assume the player is fully rational, then the (state-)entropy-regularized version of 'empowerment' and our individual power metric I fulfill the inequality

$$\begin{aligned}\mathfrak{E}^\zeta &= \max_\pi \left(\mathbb{I}_\pi(a; s) - (\zeta - 1)\mathbb{H}_\pi(s|a)\right) \\ &\leq I_h = \log_2 \sum_s \max_a \Pr(s|a)^\zeta.\end{aligned}$$

Also, I_h and $\mathfrak{E}^\zeta$ share their range $[-\log_2 k, \log_2 k]$. Let's define

$$\begin{aligned} p_{as} &= \Pr(s|a), \\ q_s &= \max_a p_{as}/Z, & Z &= \sum_s \max_a p_{as}, \\ y_s &= q_s^\zeta / Y, & Y &= \sum_s q_s^\zeta \leq 1. \end{aligned}$$

Consider any $\pi \in \Delta(\mathcal{A})$ and use the shortcuts $\pi_a = \pi(a)$ and $p_s = \sum_a \pi_a p_{as}$. Then the Kullback–Leibler divergence

$$D_{KL}(p_{a\cdot}||q) = \sum_s p_{as} \log_2 \frac{p_{as}}{q_s} \leq \sum_s p_{as} \log_2 Z = \log_2 Z$$

and thus

$$\begin{aligned}
&\mathbb{I}_\pi(a; s) - (\zeta - 1)\mathbb{H}_\pi(s|a) \\
&= \zeta \mathbb{I}_\pi(a; s) - (\zeta - 1)\mathbb{H}_\pi(s) \\
&= \zeta \sum_a \pi_a \sum_s p_{as} \log_2 \frac{p_{as}}{p_s} - (\zeta - 1)\mathbb{H}_\pi(s) \\
&= \zeta \sum_a \pi_a \sum_s p_{as} \log_2 \frac{p_{as}}{q_s}\frac{q_s}{p_s} - (\zeta - 1)\mathbb{H}_\pi(s) \\
&= \zeta \sum_a \pi_a \sum_s p_{as} \log_2 \frac{p_{as}}{q_s} - \zeta \sum_s \sum_a \pi_a p_{as} \log_2 \frac{p_s}{q_s} \\
&\qquad - (\zeta - 1)\mathbb{H}_\pi(s) \\
&= \zeta \sum_a \pi_a D_{KL}(p_{a\cdot}||q) - \zeta \sum_s p_s \log_2 \frac{p_s}{q_s} \\
&\qquad + (\zeta - 1) \sum_s p_s \log_2 p_s \\
&= \zeta \sum_a \pi_a D_{KL}(p_{a\cdot}||q) + \sum_s p_s \log_2 q_s^\zeta \\
&\qquad - \zeta \sum_s p_s \log_2 p_s \\
&\qquad + (\zeta - 1) \sum_s p_s \log_2 p_s \\
&\leq \zeta \sum_a \pi_a \log_2 Z - \sum_s p_s \log_2 \frac{p_s}{q_s^\zeta} \\
&= \zeta \log_2 Z - \sum_s p_s \log_2 \frac{p_s}{y_s Y} \\
&= \zeta \log_2 Z - D_{KL}(p||y) + \log_2 Y \\
&\leq \zeta \log_2 Z + \log_2 Y = \log_2 \sum_s (q_s Z)^\zeta = I_h
\end{aligned}$$

for all π, proving the claim. □

We remark that the proof for the case $\zeta = 1$ is similar to a derivation from Myers *et al.* [2024].

Notice that as the human becomes boundedly rational and β_h decreases, I_h will decrease and the inequality will stop holding. So, in a sense, the channel-capacity-based 'empowerment' metric corresponds to fully rational actors, while our metric I_h is sensitive to bounded rationality.

Conjecture 1. *Similar inequalities will hold in the sequential decision (MDP) case between I_h for a suitable choice of the goal set $\mathcal{G}_h$ and 'empowerment'-like metrics such as*

$$\mathfrak{E}^{\zeta}(s) = \max_{\ell \in \Delta(\mathcal{A}(s))} \Big(\mathbb{I}_{s,\ell}(a; s') - (\zeta - 1)\mathbb{H}_{s,\ell}(s'|a) + \gamma \mathop{\mathbb{E}}_{s' \sim s,\ell} \mathfrak{E}^{\zeta}(s'). \Big)$$

Despite these relationships, a fundamental difference between I_h and $\mathfrak{E}$ remains: the interpretation of the policies π occurring in their definitions. The policy π_h that I_h is based on is a function of state *and goal* g_h, has typically low entropy as it aims to attain g_h (indeed has zero entropy if h is fully rational), and might be found by standard dynamic programming or RL approaches (depending on the model of human behavior). In contrast, the maximizing "policy" π in $\mathfrak{E}$ has no real use, has typically rather high entropy in order to maximize $\mathbb{I}(a_h; s')$, and is typically harder to find.

A.2 Population-Level Human Power Metrics

Aggregation across humans: present aggregate human power $P(s)$

Proof of Proposition 4. Again, we start with the observation that this is to hold across games and any number of humans. In view of the Proposition 2, $I_h(s)$ can take thus any real value. So (P0) means we are now seeking a symmetric aggregation function $\Phi(x_1, \ldots, x_n)$ for arbitrary finite sequences of reals. Let Φ_n be its restriction on length-n sequences again.

(P1–3) then again imply separability of Φ_n with functions f_n^P, F_n^P [Debreu, 1959].

(P4) implies that $\lim_{y\to\infty} \Phi(x_1, \ldots, x_n, y) = \Phi(x_1, \ldots, x_n)$ for all real $x_1, \ldots, x_n$, which again implies that F_n^I and f_n^I are the same for all n and that $\lim_{y\to\infty} f_n^I(y) = 0$. Let us abbreviate these common f^I, F^I by just f, F.

Also (P5) can be treated very similarly, only using $\tilde{f}(y) = f(\log_2 y)$ instead of f. This is because I_h is a logarithmic quantity, so then all $C_h^{g_h}$ get multiplied by $\gamma \in (0, 1)$, I_h gets added $\delta := \zeta \log_2 \gamma < 0$.

In other words, (P5) implies: If for any real values x_i, x_i', $\sum_i f(x_i) < \sum_i f(x_i')$ and $\delta < 0$, then $\sum_i f(x_i + \delta) < \sum_i f(x_i' + \delta)$.

Since $\tilde{f}(\gamma y) = f(\log_2 \gamma + \log_2 y)$, this is equivalent to: If for any positive values y_i, y_i', $\sum_i \tilde{f}(y_i) < \sum_i \tilde{f}(y_i')$ and $\gamma \in (0, 1)$, then $\sum_i \tilde{f}(\gamma y_i) < \sum_i \tilde{f}(\gamma y_i')$.

This time, $y = 0$ is not a possible value since $y = 2^x$, hence we can only infer at this point that either

(i) $\tilde{f}(y) = \alpha(y^{\xi} + \beta)$ with $\xi > 0$, $\alpha > 0$, and some β.

(ii) $\tilde{f}(y) = -\alpha(y^{-\xi} + \beta)$ with $\xi > 0$, $\alpha > 0$, and some β.

(iii) $\tilde{f}(y) = \alpha \log y + \beta$ with $\alpha > 0$ and some β.

Now either of (P6) and (P7) implies type (ii):

(P6) is the traditional form of Pigou–Dalton this time (not the dual one of Proposition 2), so we can apply Dasgupta *et al.* [1973] directly to infer that f must be strictly concave (not convex as before). As $f(x) = \tilde{f}(2^x)$, only type (ii) fulfils this.

Alternatively, (P7) requires $f(x)$ to be bounded from above, otherwise a sufficiently large x_3 would always make the trade-off profitable. Because (i) and (iii) are unbounded, this time the only solution is (ii).

So we get $f(x) = \tilde{f}(2^x) = -\alpha(2^{-\xi x} + \beta)$.

(P4) implied that $\lim_{y\to\infty} f(y) = 0$, hence $\beta = 0$.

Regarding F, from the previous propositions and the independence of Γ^1 and Γ^2, we know that $I_h(s^1, s^2) = I_h(s^1) + I_h(s^2)$, hence (P8) implies

$$F(f(I_h(s^1) + I_h(s^2))) \tag{17}$$
$$= F(f(I_h(s^1))) + F(f(I_h(s^2))) \tag{18}$$

Because this applies across games, and $f(x) = -\alpha(2^{-\xi x} + \beta)$, we have

$$F(-\alpha 2^{-\xi x} 2^{-\xi x'}) \tag{19}$$
$$= F(-\alpha 2^{-\xi x}) + F(-\alpha 2^{-\xi x'}) \tag{20}$$

for all reals x, x'. Substituting $z = 2^{-\xi x} > 0$ and $G(z) = F(-\alpha z)$, we get

$$G(zz') = G(z) + G(z') \tag{21}$$

for all $z, z' > 0$, hence $G(z) = -\lambda \log_2 z$, where $\lambda > 0$ because F is strictly increasing. This implies

$$F(y) = G(-y/\alpha) = -\lambda \log_2 \frac{y}{-\alpha} \tag{22}$$

as claimed. □

Proof of Corollary 4.1. Assume $k \leq 2^{\xi} - 1$, $I_{h_i}(s) = 1$ for $i = 0 \ldots k$, $I_{h_0}(s') = 0$, and $I_{h'}(s') = I_{h'}(s)$ for all $h \notin \{h_0, \ldots, h_k\}$. Then $-2^{-P(s)} = -(k+1)2^{-\xi\times 1} + c \geq -2^{\xi}2^{-\xi} + c = -1 + c$ and $-2^{-P(s')} < -2^{-\xi\times 0} + c = -1 + c$, where $c = \sum_{h'} 2^{-\xi I_{h'}(s)}$ represents the remaining humans. Hence $P(s) > P(s')$ as claimed. □

Aggregation along time: trajectory-specific aggregate human power $T(s_{\geq t})$

Proof of Proposition 5. Again, we start with the observation that this is to hold across games and any number of humans. In view of the Proposition 4, $P(s)$ can thus take any real value. So we are now seeking an aggregation function $\Phi(x_1, \ldots, x_n)$ for arbitrary finite sequences of reals. This time, however, it is not symmetric, as implied by (T6). Let Φ_n be its restriction on length-n sequences again.

(T1–3) then imply separability of Φ_n with functions $f_{n,i}^P$ (now depending on position i) and F_n^P [Debreu, 1959].

Because $I_h(s) \to \infty$ for all h implies $P(s) \to \infty$, (T4) implies that $\lim_{y\to\infty} \Phi(x_1, \ldots, x_n, y) = \Phi(x_1, \ldots, x_n)$ for all $x_1, \ldots, x_n$, which again implies that F_n^I and $f_{n,i}^I$ do not

depend on n and that $\lim_{y\to\infty} f^I_{n,i}(y) = 0$. Let us abbreviate these common f^I_i, F^I by just f_i, F.

(T5) implies that there are transformations ψ_n with $\sum_{i=1}^n f_{i+1}(x_i) = \psi_n(\sum_{i=1}^n f_i(x_i))$ for all n and all $x_1, \ldots, x_n$. A classical results from decision theory shows that then there is a real value γ_r and constants c_i so that $f_{i+1}(x) = \gamma_r f_i(x) + c_i$ [Koopmans, 1960]. As the constants can be absorbed into F, we know that there are continuous and strictly increasing functions F, f and some γ_r so that

$$\Phi(x_1, \ldots, x_n) = F\big(\sum_{i=1}^n \gamma_r^i f(x_i)\big). \tag{23}$$

Strict monotonicity requires $\gamma_r > 0$.

(T6) then directly implies $\gamma_r < 1$.

Assuming (T7), notice that γ_h is any value in $(0, 1)$ (not necessarily equal to γ_r). Now it follows from [Pfanzagl, 1959] that a function of the form (23) that fulfills

$$\Phi(x_1, \ldots, x_n) < \Phi(y_1, \ldots, y_n) \tag{24}$$

$$\Leftrightarrow \Phi(x_1 + \delta, \ldots, x_n + \delta) < \Phi(y_1 + \delta, \ldots, y_n + \delta) \tag{25}$$

for all x_i, y_i from some interval and all δ from some interval must have an f of either of three claimed forms, i.e., linear, positive exponential, or negative exponential.

The outer transform F is then fixed by (T8) as in the proof of the previous proposition. □

Aggregation across uncertainty: long-term aggregate human power $L(s)$

Proof of Proposition 6. Here we are back in the von-Neumann–Morgenstern world. As $T(s_{\geq t})$ can be any real number, we seek a continuous and strictly increasing aggregation function $\Phi(\ell)$ for $\ell \in \Delta(D)$ with $D = (-\infty, \infty)$ that fulfills ordinal independence. Again defining $\ell_1 \preceq \ell_2$ as in Eqn. (9), the same proof as the one for Proposition 1 from Eqn. (9) on gives us separability: $\Phi(\ell) = F(\mathbb{E}_{x\sim\ell} f(x))$ with continuous and strictly increasing F, f.

As in the previous proof, (L4) implies that the ordering is translation invariant and the same argument as there gives the claimed three possible functional forms. □

A.3 Analysis of Paradigmatic Situations

Choosing an optimal menu size

Proof of Proposition 7. Assume the robot can choose between states s_k for all $k \geq 1$ so that $|\mathcal{A}_h(s_k)| = k$ and each $a_h \in \mathcal{A}_h(s_k)$ deterministically fulfills a separate possible goal $g_h \in \mathcal{G}_h$. Then

$$C_h^{g_h}(s_k) = \left(\frac{e^{1\times\beta_h}}{e^{1\times\beta_h} + (k-1)e^{0\times\beta_h}}\right)^\zeta,$$

$$I_h(s_k) = \log_2 k + \zeta \log_2 e^{\beta_h} - \zeta \log_2(e^{\beta_h} + (k-1)).$$

The latter is maximal for

$$k^* \approx (e^{\beta_h} - 1)/(\zeta - 1),$$

so the robot would most likely choose to got to s_{k^*} to present the human with an optimal number of options that does not overwhelm them in view of their bounded rationality. □

Asking for confirmation

Proof of Proposition 8. Let $p_k = (1-\epsilon)^k$ and $q_k = 1 - p_k - \epsilon^k$. Then the probability of getting the correct result after exactly $n+1$ rounds of being asked for A or B and then being asked for confirmation $k-1$ times is $q_k^n p_k$, which is discounted by h at factor $\gamma_h^{k(n+1)}$. Hence for each of the two goals $g_h \in \{\text{A,B}\}$,

$$C_h^{g_h}(s_k) = \gamma_h^k p_k \sum_{n=0}^{\infty} (\gamma_h^k q_k)^n = \frac{\gamma_h^k p_k}{1 - \gamma_h^k q_k},$$

and so

$$2^{I_h(s_k)} = 2\left(\frac{\gamma_h^k p_k}{1 - \gamma_h^k q_k}\right)^\zeta.$$

This is maximal where $\gamma_h^k p_k/(1 - \gamma_h^k q_k)$ is maximal. As the first-order condition is transcendental, we have to resort to numerical solutions. These show that k^* is increasing in γ_h and concave in ϵ, and that the ϵ for which k^* is largest is increasing in γ_h. For $\gamma_h \to 1$, we get $k^* \approx \ln(1-\gamma_h)/\ln\epsilon$, which is increasing in ϵ.

For $\gamma_h = 0.99$ and $\epsilon = 0.1$, the maximum is at $k^* = 2$, i.e., r will ask back once before acting. □

Making commitments

Proof of Proposition 9. Let s^n be the successor of s_0 in which r has made no commitment and let π_r^* be the $L(s^n)$-maximizing policy, i.e., what r would do after not committing. We have to prove that it is optimal to commit to π^*.

For any state s in the "not committed" subgame Γ^n starting at s^n, let $f^*(s)$ be the corresponding state in the subgame Γ^* in which r has committed to using π_r^*. (Note that some of the states $f^*(s)$ might not be reachable from $s^* = f^*(s^n)$, but that is irrelevant for the following.) By assumption, for all $g_h \in \mathcal{G}_h$, $s \in g_h$ if and only $f^*(s) \in g_h$.

For any human policy π_h^n for Γ^n, let $f^*(\pi_h^n)$ be the equivalent policy for Γ^*, $f^*(\pi_h^n)(f^*(s)) = \pi_h^n(s)$. Denote the two respective goal-dependent value functions of h in Γ^n and Γ^* by $V^n_{g_h,\pi_h^n}$ and $V^*_{g_h,f^*(\pi_h^n)}$. Note that because there are no other humans and because r behaves equivalently in Γ^n and Γ^* (because it has only committed to what it would anyway do), we have $V^*_{g_h,f^*(\pi_h^n)}(f^*(s)) = V^n_{g_h,\pi_h^n}(s)$ for all g_h, π_h^n and $s \in \mathcal{S}^n = \mathcal{S}|_{\Gamma^n}$.

Now fix some g_h and let π_h^n and π_h^* be h's optimal policies for attaining g_h in Γ^n and Γ^* given h's beliefs π_r^0 about r's behavior in the respective subgames. Since in Γ^*, r has bindingly committed to using π^* and h only attaches probability to actions r has not ruled out, h correctly believes that r is using π^* in Γ^*. So π_h^* is the $V^*_{g_h}$-maximizing policy for Γ^* under π^*. In particular, $V^*_{g_h,f^*(\pi_h^n)} \leq V^*_{g_h,\pi_h^*}$. Generically, if different robot actions lead to different consequences so that h's optimal policy depends on their beliefs about r's actions, then knowing the true π_r would strictly increase goal-attainment and the above inequality would be strict.

Also, if r committed to anything else than π_r^*, leading to some successor state $s^\dagger$ starting a subgame $\Gamma^\dagger$ in which h's optimal policy was $\pi^\dagger$, then in the same fashion we see that

$V^*_{g_h,f^\dagger(\pi^\dagger_h)} \leq V^*_{g_h,\pi^*_h}$. In this, $f^\dagger$ would be the corresponding state mapping from $\Gamma^\dagger$ to Γ^*. Again, in the generic case the inequality would be strict.

Now we have

$$\begin{aligned} C^{g_h}_h(s) &= V^n_{g_h,\pi^n_h}(s) \\ &= V^*_{g_h,f^*(\pi^n_h)}(f^*(s)) \\ &\leq V^*_{g_h,\pi^*_h}(f^*(s)) \\ &= C^{g_h}_h(f^*(s)) \end{aligned}$$

for all $s \in \mathcal{S}^n$, and similarly for all $s \in \mathcal{S}|_{\Gamma^\dagger}$.

Since all aggregations are strictly increasing and r is optimizing, we also have $L(s^*) \geq L(s^n), L(s^\dagger)$ (and generically strictly so) by induction.

Hence it is weakly (and generically strictly) optimal for r to commit to π^*_r rather than to any other or no policy at all. □

Of course, communicating all details of a complicated policy π^*_r to h will in general not be possible, so r will in general have to decide how exactly to use its limited communication possibilities. This is what the following example is about.

Several buttons: k-means clustering of goals and policies Another insightful example is where in addition to the initial commitment stage for r, there is a subsequent choice by h before the actual game Γ is played. To exemplify this, assume r has $k > 1$ many buttons each of which it can label in s_c with one of its own policies for Γ, after which h can press one of the buttons, committing r to the respective policy, and then Γ is played with the committed policy. If $|\mathcal{H}| = 1$ and $\beta_r = \beta_h = \infty$ as before, and if $\gamma_r \ll 1$ so that r only cares for h's immediate power, then r would aim to find that set of policies π^i_r, $i = 1 \ldots k$, that covers h's goal space best in the sense that it maximizes $L(s_c)$ as computed on the basis of $C^{g_h}_h(s) = \max^k_{i=1} C^{g_h}_h(s|\pi^i_r)$ because, depending on g_h, h would press the button for that policy π^i_r which maximizes its capability to maximize the probability to attain g_h. Finding the best-covering k policies might however not be possible exactly due to the high dimension of the policy space.

A natural approximation would be to use a variant of k-means clustering such as the following in order to partition $\mathcal{G}_h$ into k sets $\mathcal{G}^i_h$ and identifying the corresponding optimal robot policies π^i_r. Start with k randomly selected goals g^i_h and put $\mathcal{G}^i_h = \{g^i_h\}$. Then alternate the following two steps. For each $\mathcal{G}^i_h$, estimate the optimal π^i_r as usual (using backward induction or reinforcement learning, just with a goal set restricted to $\mathcal{G}^i_h$). Then, for each $g_h \in \mathcal{G}_h$, compute $C^{g_h}_h(s_c|\pi^i_r)$ for all i, put $j = \arg\max_i C^{g_h}_h(s_c|\pi^i_r)$, and assign g_h to $\mathcal{G}^j_h$. Alternate until (approximate) convergence.

Preventing destruction but allowing being paused

Proof of Proposition 10. We assume $\rho = 1$ and that Γ^2 has the following states, actions, and transitions:

- s_d The robot is destroyed. r and h can only pass, the successor state is again s_d.
- s_{p1} r is paused, P is enabled, D not. h might or might not toggle P.
- s_{p2} r is paused, P and D are enabled. h might or might not toggle P and might or might not press D (4 combinations).
- s_{a0} r is active (neither paused nor destroyed), both buttons disabled. r might enable P or both P and D.
- s_{a1} r is active, P is enabled, D not. h might press P. If they don't, r might disable P or enable D.
- s_{a2} r is active, P and D enabled. h might press either. If they don't, r might disable D or both P and D.

So in the full game, we have $I_h(s_d|s_{p1}|s_{p2}|s_{a0}|s_{a1}|s_{a2}) = x|x+1|x+2|y|y+1|y+2$ because the only interaction between Γ^1 and Γ^2 is due to whether r is active or not. Because $\beta_r = \infty$, r chooses either to disable both P and D, or enable only P, or enable both, whenever they get the chance. If it is optimal to disable both once, it is optimal to keep them disabled; let us denote the resulting $-2^{-L(s_{a0})}$ by W_0. If it is optimal to enable only P, that keeps being optimal; let us denote the resulting $-2^{-L(s_{a1})}$ by W_1. If it is optimal to enable both, that keeps being optimal; let us denote the resulting $-2^{-L(s_{a2})}$ by W_2. Let us put $\alpha = -\xi\eta < 0$, $\gamma = \gamma_r$, $\epsilon = 1-\gamma$, $\delta = 1-\gamma p$, $A_0 = -y^\alpha < A_1 = -(y+1)^\alpha < A_2 = -(y+2)^\alpha$, $B_1 = -(x+1)^\alpha < B_2 = -(x+2)^\alpha$, $C = -x^\alpha/\epsilon$.

Then solving the system of linear equations for the Markov chains of Γ^2 states resulting from the three possible robot policies gives

$$W_0 = A_0/\epsilon, \tag{26}$$

$$W_1 = \frac{\delta A_1 + \gamma p B_1}{\epsilon}, \tag{27}$$

$$W_2 = \frac{\delta A_2 + \gamma p B_2 + \gamma q C}{1-\gamma(1-q)} \tag{28}$$

If both $x, y \gg 1$ and $d = y - x$, we have approximately

$$W_1 > W_0 \Leftrightarrow p\gamma d < 1, \tag{29}$$

$$W_2 > W_0 \Leftrightarrow q\gamma d < \epsilon(2-\gamma p d), \tag{30}$$

$$W_2 > W_1 \Leftrightarrow q\gamma(1+d\delta) < \epsilon. \tag{31}$$

Hence, enabling both buttons is optimal only if

$$q\gamma < \epsilon \min\{(2/d - \gamma p), 1/(1+d\delta)\},$$

and disabling both is optimal only if

$$\gamma d > \max\{1/p, \epsilon(2-\gamma p d)/q\}.$$

For moderate x, y, we can derive these approximations:

If $q \sim p \ll \epsilon$ or $q \ll p \sim \epsilon$, $W_i \approx A_i/\epsilon$, hence $W_2 > W_1 > W_0$, hence keeping both buttons enabled is optimal.

If $p \ll q \sim \epsilon \ll 1$ or $p \sim q \sim \epsilon \ll 1$, then W_2 is largest iff $q/\epsilon < (A_2 - A_1)/(x^\alpha + A_1)$, otherwise W_1 is largest; so r will keep both buttons enabled if q is small enough, else only the pause button.

If $\epsilon \ll q$, then W_1 is largest, so only the pause button will be enabled.

If $p \approx 1$, then W_0 is largest (unless $x + 1 > y$, in which case W_1 is still largest), so both buttons will be disabled.

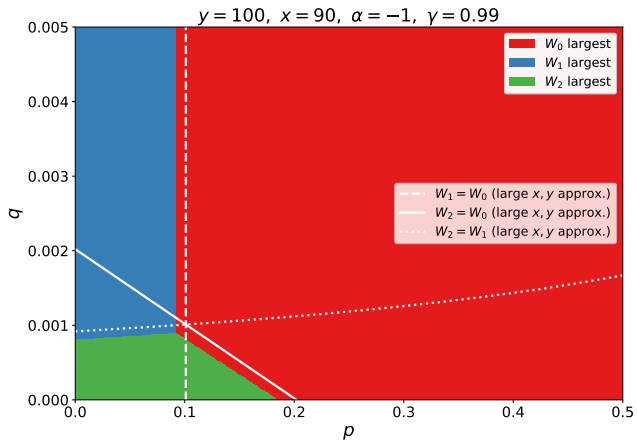

Figure 2: Optimal robot action in the pause and destroy buttons example. W_0: disable both buttons. W_1: enable only the pause button. W_2: enable both buttons. White dotted/dashed lines: boundaries from large x, y approximation.

For moderate p, q, ϵ, the analysis is more involved. Numerical simulation for $y = 100$, $x = 90$, $\gamma = 0.99$, and varying p, q shows that if q is non-negligible and approximately $p < 1/11$, the robot will enable the pause button but not the destroy button (Fig. 2). □

B Example of a Boundedly Rational Human Behavior Model

We might equip r with a simple model of h's decision making that focuses on giving r the right incentives. It assumes h cannot realize the maximal goal attainment probability due to a variety of reasons relating to exploration, imperfect action implementation, information constraints, others' behavior, and potentially state-dependent cognitive limitations. In particular, r would *not* assume h to have correct beliefs about others' behavior that would lead to an equilibrium (Nash, quantal response, etc.). Instead, r would model h as having beliefs $\mu_{-h}(s, g_h)$ about other humans' behavior (where '$-h$' is short for $\mathcal{H} \setminus \{h\}$). These would also reflect social norms (which LLM-based systems already understand, [Smith *et al.*, 2024]). Hence r would assume goal-dependent state-action values $Q_h(s, g_h, a_h)$ that are based on $\mu_{-h}(s, g_h)$.

The robot could assume that h uses a mixture (governed by a probability ν_h) between (i) *habitual, 'system-1'* behavior encoded in some default policy $\pi_h^0(s, g_h)$, again reflecting social norms, and (ii) *boundedly rational, 'system-2'* behavior represented by a Boltzmann policy with rationality parameter β_h. As in Ghosal *et al.* [2023], β_h may be state-dependent, which would give r incentives to choose states with larger β_h, and β_h might be estimated from observations [Safari *et al.*, 2024]. Reflecting social norms, ν_h, μ_{-h}, and π_h^0 are the only significantly "semantically loaded" elements of the world model.

To model *human beliefs about the robot's actions,* the world model should contain information about what actions r has previously committed to choose from in a state s: the action set $\mathcal{A}_r(s)$ should only contain those actions, and different commitment histories should be considered different states.

C Choice of reliability preference ζ

Assume a multi-armed bandit problem where $\mathcal{G}_h$ is a partition of the outcome set into k blocks. Then we have seen that $I_h \leq \log_2 k$ regardless of ζ.

(i) If for each block of outcomes, h has an action bringing it about for sure, then $I_h = \log_2 k$. This case might be interpreted as h being all powerful and "nature" being powerless (as the outcome is exclusively determined by h).

(ii) If the same outcome obtains for sure regardless of what h does, we have $I_h = \log_2(1 + (k-1) \times 0) = 0$ regardless of ζ. This case might be interpreted as neither h nor "nature" having any power.

(iii) In the worst case, each g_h is reached with probability $1/k$ regardless of what h does. In that case, $I_h = \log_2 k(1/k)^\zeta = (1-\zeta)\log_2 k$.

The latter case might be interpreted as h being powerless but "nature" being all powerful. This interpretative symmetry suggests that in this case I_h should be as far from 0 as it is in case (i), i.e., $I_h = -\log_2 k$. This is the case if and only if $\zeta = 2$.

Another argument for the choice $\zeta = 2$ is that the corresponding entropy-regularized channel capacity then has a particularly simple form:

$$\mathfrak{E}^2 = \max_\pi \big(\mathbb{I}_\pi(a; s) - \mathbb{H}_\pi(s|a)\big).$$

Finally, I_h with $\zeta = 2$ is also closely related to collision entropy (Renyi-entropy of order 2).

D Hedging against the robot becoming defunct or corrupted

Such a hedging could be achieved in several ways.

One can include a rate $\delta > 1 - \gamma_r$ of the robot becoming temporarily or permanently *defunct* and "passes" on each step. To achieve this, wrap a learned base world model into a wrapper model that adds this transition. This should prevent policies that make humans depend on the robot's presence too much.

One can also include a flag "robot corrupted" into the state space of the wrapped world model and add a positive rate δ' of becoming permanently *corrupt* into the transition kernel. Then, when calculating $L(s)$ recursively on the basis of $L(s')$ (which requires $\rho = 1$ however), multiply $L(s')$ by -1 if corruptness$(s') \neq$ corruptness(s), and when calculating $P(s)$, multiply it by -1 if corruptness$(s) = 1$.

E Alternative metrics

E.1 Alternative aggregation order

Instead of aggregating $I_h \to P \to T$, one could instead first aggregate I_h into *individual lifetime power* L_h and then into

aggregate long-term human power L' via

$$L_h(s_t) = -\log_2 \mathop{\mathbb{E}}_{s_{\geq t} \sim s_t, \pi} \big(\textstyle\sum_{u \geq t} \gamma_{rh}^{u-t} 2^{-\eta I_h(s_u)} \big)^{\rho}, \quad (32)$$

$$L'(s) = -\log_2 \textstyle\sum_{h \in \mathcal{H}} 2^{-\xi L h(s)}, \quad (33)$$

where γ_{rh} represents the robot's relative weighing of h's future power vs h's current power, which could be informed by h's life expectancy.

We have chosen the first ordering mostly for four pragmatic reasons: It gives us a metric of present aggregate human power P that let's one reason about intergenerational and humanity-vs-AI power balances. Also, in view of equation (7), its transformed version $-2^{-\eta P}$ can be interpreted as an "intrinsic reward" for the robot that opens up the possibility of applying some form of model-based reinforcement learning to calculate π_r.

Third, instead of tasking the robot to softly maximize L, a combination of P and L could be used as a constraint within which the robot aims to assist its current principal, say via cooperative inverse reinforcement learning. One version of such a permissibility constraint could be to demand that the robot prevents aggregate human power from decreasing in the long run. More precisely, a_r would be *permissible* in s iff

$$\mathop{\mathbb{E}}_{s' \sim s, a_r, \pi_{\mathcal{H}}} (-2^{-L(s')}) \geq -2^{-\eta P(s)}/(1-\gamma_r), \quad (34)$$

where the RHS is the expected "value" of a constant-aggregate-power trajectory at the current generation's power level.

Finally, the chosen ordering (humans, then time) seems to be more common in applied welfare economics in long-term contexts across generations, e.g. in discounted-utilitarian evaluation in growth and climate economics [Ramsey, 1928; Nordhaus, 2008], while the opposite ordering (time, then humans) is more popular in welfare philosophy.

E.2 Avoiding independent sets of possible goals

The necessity to specify a set of possible goals $\mathcal{G}_h$ leads to potential distortions through misspecification and manipulation, e.g. because some goals are individuated in much more variations than others or some goals are much broader than others.

This motivates considering to strip $\mathcal{G}_h$ of its independence and rather make it a "canonical" function of the world model's state space $\mathcal{S}$.

In the case of a tree-shaped transition graph, a particularly convenient choice is to simply make each possible state a possible goal, so that

$$\mathcal{G}_h = \{\{s\} : s \in \mathcal{S}\}. \quad (35)$$

If we assume in addition that the goal-conditioned human policy $\pi_h(s, g_h)$ decomposes into stepwise policies $\pi_h(s, s')$ where s' is the successor of s in direction of the goal state g_h, then computation of I_h simplifies to a recursive formula (a kind of nonlinear Bellman equation as $\zeta > 1$):

$$2^{I_h(s)} = 1 + \gamma_h^{\zeta} \sum_{s' \in \mathcal{S}} \Pr(s' \mid s, \pi_h(s, s'), \pi_{-h}, \pi_r)^{\zeta}\, 2^{I_h(s')}.$$

A convenient side-effect is then that the computational complexity declines because no summation over g_h is necessary any longer.

In the case of non-tree-shaped transition graphs, the same formula can be used. In that case the interpretation is that *each possible goal is a finite continuation of the state trajectory starting at s,* and $\pi_h(s, s')$ is the policy the world model assumes that h uses if they want to get to s' next when in s.