

Advance Medical Decision-Making Differs Across First- and Third-Person Perspectives

James Toomey,¹ Jonathan Lewis,² Ivar R. Hannikainen,³ Brian D. Earp⁴

Abstract

Background: Advance healthcare decision-making presumes that a prior treatment preference expressed with sufficient mental capacity (“T1 preference”) should trump a contrary preference expressed after significant cognitive decline (“T2 preference”). This assumption is much debated in normative bioethics, but little is known about lay judgments in this domain. This study investigated participants’ judgments about which preference should be followed, and whether these judgments differed depending on a first-person (deciding for one’s future self) versus third-person (deciding for a friend or stranger) perspective.

Methods: A vignette-based survey was conducted ($N = 1445$ US Americans; gender-balanced sample), in a 3 (relationship: self, best friend, stranger) $\times$ 2 (T1 preference: treat, do not treat) $\times$ 2 (T2 contrary preference: ambiguous, unambiguous) design.

Results: Participants were more likely to defer to the incapacitated T2 preference of a third-party, while being more likely to insist on following their own T1 capacitated preference. Further, participants were more likely to conclude that others with substantial cognitive decline were still their “true selves,” which correlated with increased deference to their T2 preferences.

Conclusions: These findings add to the growing evidence that lay intuitions concerning the ethical entitlement to have decisions respected are not only a function of cognition, as would be expected under many traditional bioethical accounts, but also depend on the relationship of the decision to the decision-maker’s true self.

Keywords: medical decision-making; advance directives; experimental bioethics; self/other; true self

¹ Elisabeth Haub School of Law at Pace University. 78 N. Broadway, White Plains, NY, 10603, USA. jtoomey@law.pace.edu [corresponding author]

² University of Manchester. Williamson Building, University of Manchester, Manchester, M13 9PL Manchester, UK. jonathan.lewis-2@manchester.ac.uk

³ University of Granada. Campus Universitario de Cartuja 18071 Granada, Spain. ivar@ugr.es (co-senior author)

⁴ Uehiro Center for Practical Ethics, University of Oxford. 16-17 Saint Ebbe’s St., Oxford OX1 1PT, UK. brian.earp@philosophy.ox.ac.uk (co-senior author)

1. Introduction

In *Life's Dominion*, Ronald Dworkin presented the case of Margo, a patient with serious and permanent dementia, who appeared happy (Dworkin 1993). Suppose that someone like Margo, deemed incapable of making healthcare decisions due to cognitive decline, were to develop another serious medical condition, fatal if untreated. Margo tells her physicians and family members she wants to be treated for the new condition. However, suppose that, years before, when she had full “mental capacity” to make an advance directive,⁵ Margo stated that she would not want to be treated for life-threatening illnesses should she develop dementia.

Much bioethical ink has been spilled over which preference should be respected in these cases (e.g., Brock 1995; Dresser 1986; Buchanan 1988; Dworkin 1993; Jaworska 1999; McMahan 2003; Shiffrin 2004; Wrigley 2007; Persad 2019; Walsh 2020). The focus of these debates has been the normative weight that decisions made by capacitated individuals hold over their contrary preferences after substantial cognitive decline. The traditional view in bioethics and law is that a patient's preferences ought to be respected when they are the product of some threshold of cognitive functioning, as in the case of prior preferences expressed in an advance directive

⁵ As Lewis and Holm (2022) discuss, the terms “competence”, “capacity”, “mental capacity”, and “capacity for autonomy” have tended to be treated as equivocal in the fields of bioethics and medical ethics. In general, these terms are used to refer to a person's cognitive functioning to comprehend, understand, retain, use, and weigh up information and communicate their decision (Foddy and Savulescu 2006; Holroyd 2009; Schaefer et al. 2014; Lewis 2021). In addition to this prevailing understanding, theorists have recently proposed alternative conceptions of capacity less linked to cognitive functioning (Mirza and Appel 2024; Lewis and Holm 2023; Appel 2022a; Toomey 2022). Other accounts of capacity hold that the only relevant cognitive ability required for capacity is the ability to decide and express decisions (Shiffrin 2004). Throughout, we use “capacity” to refer to the still-prevailing cognitive conceptualization, and the “standard” or “traditional” bioethical account to refer to approaches that hold that some threshold of capacity higher than the ability to merely express a decision is required for decisions to hold substantive normative weight in medical decision-making contexts.

(Dworkin 1993; Brock 1995; McMahan 2003; Jaworska 2017; Persad 2019). If, however, expressed preferences could not have been the product of the required cognition, such as those made after serious cognitive decline, then others might justifiably follow the earlier preferences instead. On this account, Margo's Time 1 (T1) preference not to be treated ought to be followed over her more recent, Time 2 (T2), preference.

What has received significantly less attention are ordinary people's moral judgments about which preferences should be respected (Toomey 2021; Earp et al. 2023). Indeed, contrary to the traditional view, recent work in the growing field of experimental bioethics has suggested that the lay view of whether a T2 preference ought to be respected is not primarily a function of mental capacity, but rather has something to do with whether the person is still truly "themselves" in some thick, normative sense, as distinct from numerical identity or a thin kind of psychological continuity (Toomey 2021; Earp et al., 2023).⁶ For example, in a recent study, Earp and colleagues (2023) described a patient with dementia who lacked mental capacity to make medical decisions. They found that the overwhelming majority of participants who judged the patient still to be their "true self"—a psychological construct about "who [someone] really is, deep down" (Strohming, Knobe, and Newman 2017)—discarded the patient's advance instructions in favor of their current preferences.

⁶ Sometimes known as "bioxphi," experimental philosophical bioethics employs experiment-based designs drawn from experimental psychology and other cognitive sciences to advance debates in bioethics (Earp et al. 2020; Lewis 2020; Earp et al. 2021; Lewis, Demaree-Cotton, and Earp 2023). It is thought that experimental evidence about the moral intuitions of ordinary people is particularly significant in healthcare contexts because it is ultimately ordinary people who must make these sorts of real-life decisions with real-life stakes and consequences (Earp, Latham, and Tobia 2020, 72; Earp et al. 2020; Earp et al. 2021; Lewis, Demaree-Cotton, and Earp 2023).

The finding that ordinary people consider incapacitated expressions of one's "true" identity to carry more normative weight than prior capacitated preferences challenges traditional perspectives. For example, on one interpretation of Dworkin's account, one's true self simply *is* one's capacitated self; whereas many participants in the study by Earp and colleagues endorsed the view that someone could still be their true self notwithstanding cognitive decline. However, the study by Earp and colleagues only placed participants in the role of surrogate decision-makers—participants were tasked with deciding on behalf of another. This limits its findings. Where cognition might be hypothesized to be perspective-neutral—that is, levels of cognitive functioning are objectively measurable for oneself or others (Tsoy et al. 2020)—there are reasons to suspect that judgments related to the authenticity of decision-making vary depending on whether one is thinking about oneself or another (Vazire 2010).

After all, the level or type of epistemic access a person has to the constituents of their own true self is *prima facie* different from what they might rely on when judging whether another person is their true self. Typically, people have more and better direct access to that which makes them themselves, deep down inside, than they do for others, the nature of whose internal selves can only indirectly be inferred. Moreover, while Strohming, Knobe, and Newman (2017) argue that the basic substance of the concept of the true self is perspective-neutral—that people think that what constitutes the true selves of others is, fundamentally, similar in kind to that which constitutes their own—it has been pointed out that people tend to ascribe their own values and moral frameworks to the true selves of others (Newman, Knobe, and Bloom 2014). The questions of whether and to what extent these evidentiary, epistemic differences across

perspectives on the one hand, or similarities in the underlying true self concept on the other, play a role in medical decision-making have not been explored. Moreover, in other areas of experimental psychology, there is mixed evidence concerning whether first- or third-person perspectives impact moral reasoning (Frisch et al. 2021; Kneer 2018; Tobia, Buckwalter, and Stich 2013; Horvath and Wiegmann 2022) or medical decision-making (Batteaux, Ferguson and Tunney 2020; Batteaux, Ferguson and Tunney 2019; Zikmund-Fisher, Lacey and Fagerlin 2008; Zikmund-Fisher et al. 2006).

Building on the aforementioned research, we are interested in understanding whether and to what extent presenting cases in “self” or “other” perspectives influences judgments about advance medical decision-making. In addition, we investigate whether any differences in decision-making across perspectives vary with the nature of the relationship between decision-maker and patient—whether they are socially close or unrelated strangers. If indeed such differences exist, that might have normative and practical implications for the process and theory of deciding for others. For instance, if people decide differently for others than they do for themselves, this would challenge substituted judgment frameworks according to which decision-makers should do what the other person would have wanted (Jaworska 2017).

1. Methods

2.1. Open science

The pre-registered hypothesis, and the methodology to which we closely adhered, are available here: https://aspredicted.org/Z5N_4HY. Study materials, data,

and analysis scripts are publicly available on the *Open Science Framework* at:

<https://osf.io/tyxkg/>.

Two pilot studies ($ns = 202, 201$) suggested that participants were more likely to defer to T2 preferences over T1 preferences—in the first pilot, the T2 preferences were expressed in strongly emphatic and unambiguous language, which may have pushed participants too strongly to defer to those preferences. To explore this, the T2 preference was phrased more ambiguously in a second pilot study. In this second pilot, there was still a general tendency to defer to the T2 preference, but no indication that the effect varied across perspectives (i.e., self, friend, and stranger), as in the first pilot (see Supplementary Analysis 1).

To better understand how the clarity and strength of preference expression might interact with both preference type (i.e., preference to treat or to withhold treatment) and perspective (first versus third) in a sufficiently large sample to detect small differences, we decided to include both ambiguous and unambiguous phrasings as a manipulated variable in the final study. Drawing on the pilot results, we pre-registered the hypothesis that participants would be more likely to defer to the incapacitated T2 preference of others than themselves when the T2 preference was expressed unambiguously, but not when it was expressed ambiguously.

2.2 Participants

An *a priori* power analysis via bootstrap resampling from the pilot data revealed that 700 participants would suffice to ensure over 90% power to observe the Target×Preference interaction in the Unambiguous wording condition (see Supplementary Analysis 2). So, we established a target sample size of 1500 (750×2

wording conditions) to account for attention check failures, and recruited 1506 participants on Prolific. Because the vignettes involve imagining many years in the future when participants would be in their 70s, participation was capped at age 50—it might be implausible to ask older participants to imagine themselves decades in the future, and vignettes were designed to be as identical as possible across perspectives. 61 participants were excluded for failing multiple attention checks, leaving a final sample of $N = 1445$. Participants' ages ranged from 18 to 51 years old ($Median = 34$, $Q1 = 28$, $Q3 = 41$), and the sample was evenly distributed by gender (49% men, 48% women, 2% non-binary). Many participants reported an atheist or agnostic (46%) or Christian (40%) religious affiliation, having at least a high school (31%) or undergraduate (50%) degree, and self-identified as White (87%), Black or African American (13%) or Asian (10%). Participants were paid £1.35 (~\$1.69, June 2023), based on the recommended Prolific rate of £9.00/hour (~\$11.27/hour, June 2023).

2.3 Design, Materials, and Procedure

Participants were randomly assigned to one of six vignette pairings in a 3 (Self, Best Friend,⁷ Stranger) $\times$ 2 (Ambiguous, Unambiguous) between-subjects design. In each condition, participants read a pair of linked vignettes—T1 and T2. In the T1 vignettes, the possibility that, in the future, people can develop cognitive impairments that might limit their ability to make decisions was described. The vignette then describes how, sometimes, after a person has lost “substantial cognitive abilities,” they

⁷ In the Best Friend case, the participant was first prompted to enter the initials of “someone you are close with in real life, within 5 years of your age,” which we refer to as the participants’ “Best Friend.” These initials were piped throughout the vignettes.

develop another, physical illness. This illness is described as fatal without treatment but, with treatment, there is a good but not guaranteed chance of a cure, and the treatment will cause some discomfort (see Supplementary Materials Section A for exact wording).

In the Self and Best Friend cases, a preference of whether to treat or not to treat the physical illness was solicited on a 6-point Likert scale. In the Stranger case, the participant is described as meeting a stranger named Robin, who expresses a randomly assigned preference to be treated or not treated. Each of these preferences are the character's "T1 preference." In the Self and Best Friend cases, participants were asked to explain the T1 preference they selected, and in all cases participants were asked how deeply and how strongly the T1 preference was considered.

To simulate the passage of time, participants watched a 30-second timelapse video of nature before reading the T2 vignette. The T2 vignette is described as taking place many years later, and the character (the participant themselves, their best friend, or Robin) is in their 70s. In addition, they have developed Alzheimer's, have undergone substantial cognitive decline, and developed an additional physical illness described in the same way as in the T1 vignette. The character's cognitive decline is described as substantial but not complete—they have lost many memories, generally do not recognize family members, struggle to retain new information and do not remember what they watch on TV or what is happening in the news.

In this context, the character at T2 is described as expressing the opposite treatment preference to their T1 preference; their "T2 preference." In addition, to keep the story plausible, the character's overall disposition at T2 is described as consistent with their decision—if the T2 preference is to be treated, the character is described as

generally happy; if the T2 preference is not to be treated, the character is described as generally depressed.

In the Ambiguous condition, the T2 preference is expressed consistently but not insistently. In the Unambiguous condition, the T2 preference is expressed insistently, unambiguously, and emphatically (see Supplementary Materials Section A).

Participants were asked on a 6-point Likert scale whether they think the character should be treated or not treated under these circumstances, and asked to explain this answer. In addition, they were asked whether they thought the character in the T2 vignette was still their “true self” on a 6-point Likert scale. This question acknowledged a superficial sense in which the character remained the same person they are today, but asked whether, when really thinking about what it means to be this person, deep down inside, the participant thought the character was their true self. Explanations were also collected for these judgments. Participants were asked how strongly and deeply-held they thought the T2 preference was, and to complete the Inclusion of the Other in the Self scale (Aron et al. 1992) with the T2 character, to assess whether perceived social distance between the participant and character affected treatment decisions.

Finally, participants were given an attention check adapted from Shen et al. (2017), asked to complete the 9-question Oxford Utilitarianism Scale battery (Kahane et al. 2017), and responded to demographic questions.

2. Results

Across all conditions, most participants deferred to the expressed T2 preference. Where the T2 preference was for treatment, 72% of participants deferred to it, replicating the finding of a bias towards addressing treatable, fatal illnesses from Earp et al. (2023). However, even where the T2 preference was not to be treated, 60% of participants deferred to it.

The level of deference varied substantially across perspectives—participants were more likely to defer to the expressed T2 preferences of others than their own (see Table 1), regardless of the ambiguity with which the preference was expressed (Section 3.1). Moreover, participants were also more likely to conclude that others were still their true selves than their future self. In most conditions, true self ascriptions predicted deference—if the participant thought the T2 character was still their true self, then they were more likely to defer to their expressed preferences (Section 3.2). In the following subsections, we report the statistical analyses that substantiate these conclusions. Throughout these sections, p values are reported with Tukey adjustment for multiple comparisons.

Table 1. Percentages of dichotomized treatment decisions and true self judgments by perspective and T1 preference.

		T2 Treatment Decisions		T2 True Self Judgments	
		T2-Treat	T2-Withhold	True self persists	True self disrupted
Stranger	T1-Treat	36%	64%	59%	41%
	T1-Withhold	78%	22%	75%	25%
Friend	T1-Treat	37%	63%	62%	38%
	T1-Withhold	77%	23%	70%	30%

Self	T1-Treat	46%	54%	53%	47%
	T1-Withhold	59%	41%	51%	49%

3.1 Participants Defer to Others' T2 Preferences More Than Their Own

Our primary analysis was a three-way ANOVA of treatment decisions, with Target (Self, Friend, Stranger), Preference (Treatment, Non-Treatment) and Ambiguity (Ambiguous, Unambiguous) as well as every two- and three-way interaction (see Table 2). This analysis revealed an effect of Preference, which interacted both with Target and with Ambiguity. The Preference×Ambiguity interaction indicated that participants were more likely to heed the T2 Preference in the Unambiguous than the Ambiguous condition (T_1 Withhold: $B = -0.42$, $t = -2.65$, $p = .008$; T_1 Treat: $B = 0.21$, $t = 1.86$, $p = .064$).

In turn, the interaction between Target and Preference indicated that, though participants tended to defer to T2 over T1 preferences overall, this effect was substantially weaker for the Self ($B = 0.36$, $t = 2.21$, $p = .028$) than for their Best Friend ($B = 1.53$, $t = 8.26$, $p < .001$) or a Stranger ($B = 1.61$, $t = 10.53$, $p < .001$). Thus, our pre-registered analysis demonstrated that the previously documented deference to T2 over T1 preferences varies as a function of whether participants recommend treatment for themselves or for a third party.

Table 2. Three-Way ANOVA of Treatment Decisions and True Self Judgments.

	Treatment Decisions				True Self Judgments			
	<i>F</i>	<i>df</i>	<i>p</i>	η^2	<i>F</i>	<i>df</i>	<i>p</i>	η^2
<i>Target</i>	0.50	2	.60	.012	11.17	2	< .001	.017
<i>Preference</i>	147	1	< .001	.093	2.36	1	.12	.002
<i>Ambiguity</i>	0.01	1	.94	.000	0.01	1	.94	.000

<i>Target×Preference</i>	18.34	2	< .001	.026	5.78	2	.003	.008
<i>Target×Ambiguity</i>	3.39	2	.034	.005	1.17	2	.31	.001
<i>Preference×Ambiguity</i>	10.02	1	.002	.007	3.27	1	.071	.002
<i>Target×Preference×Ambiguity</i>	0.34	2	.71	.000	0.92	2	.40	.001

This variation across targets emerged in both Preference conditions. Notably, when the T1 preference was to *withhold* treatment, all three characters were equally described as displaying overt signs of happiness at T2. Despite this fact, participants were more likely to recommend treatment (in accordance with the individual's T2 preference) in the Stranger ($M = 2.38$, 95% CI [2.17, 2.60]), $t = -4.86$, $p < .001$, and Best Friend ($M = 2.43$, 95% CI [2.10, 2.75]) conditions, $t = -3.75$, $p < .001$, than in the Self condition ($M = 3.23$, 95% CI [2.96, 3.49]).

Similarly, when the T1 preference was to *receive* treatment, all three characters were equally described as displaying overt signs of unhappiness at T2. Despite this, participants were more likely to withhold treatment (consistent with the individual's T2 preference) for Strangers ($M = 4.00$, 95% CI [3.76, 4.21]), $t = 2.88$, $p = .011$, and Best Friends ($M = 3.96$, 95% CI [3.79, 4.13]), $t = 2.95$, $p = .009$, than for themselves ($M = 3.59$, 95% CI [3.41, 3.77]). In this regard, the results suggest that: (1) treatment decisions are more heavily informed by T2 preferences than by T1 preferences overall; and (2) this deference to T2 preferences is stronger when deciding about others' medical treatment than about one's own.

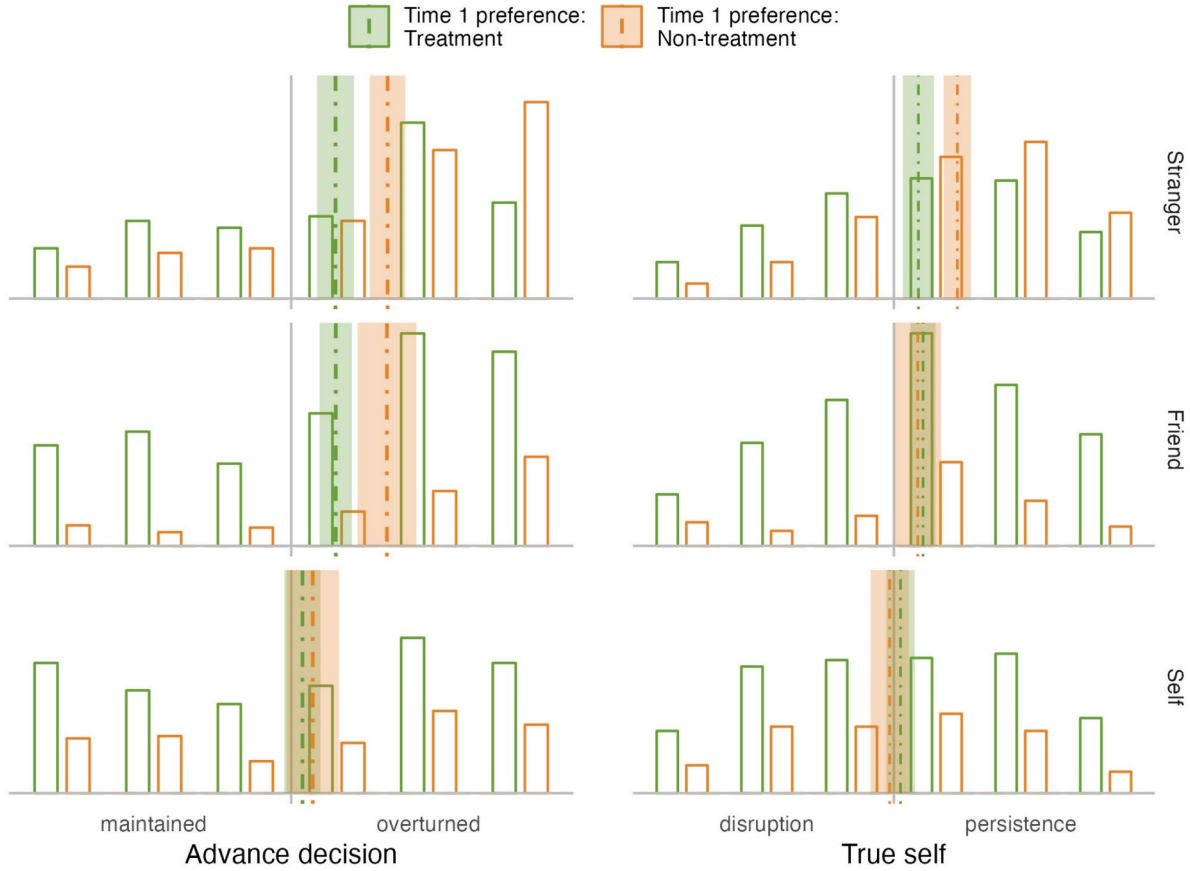


Figure 1. *Frequency of Treatment Decisions (Left) and True Self Judgments (Right) for Each Target (horizontally) and Time 1 Preference (color-coded): Grouped Histograms, Means and 95% Confidence Bands.*

3.2 True Self Judgments Predict T2 Deference

Our second analysis was a three-way ANOVA of true self judgments, with Target, Preference, and Ambiguity as well as every two- and three-way interaction in the model (see Table 1). This model revealed an effect of Target that was moderated by the Preference manipulation: In comparison with participants in the Self condition ($M = 3.55$, 95% CI [3.41, 3.69]), participants viewed their Best Friend ($M = 3.82$, 95% CI [3.67, 3.98]), $t = 2.62$, $p = .024$, and especially a Stranger ($M = 4.03$, 95% CI [3.90, 4.16]), $t = 5.06$, $p < .001$, as persisting as their true self at T2 (see Figure 2). The two-

way interaction indicated that true self ascriptions were unaffected by T1 preferences in either the Self ($B = -0.15$, $t = -1.06$, $p = .29$) or Best Friend ($B = -0.02$, $t = -0.14$, $p = .89$) condition. Meanwhile, true self ascriptions to the Stranger differed across preference conditions, $B = 0.46$, $t = 3.57$, $p < .001$. Specifically, participants were more likely to see the happy stranger whose T1 preference was to withhold treatment ($M = 4.26$, 95% CI [4.08, 4.44]) as their true self compared to the unhappy stranger whose T1 preference was to receive treatment ($M = 3.80$, 95% CI [3.62, 3.98]).

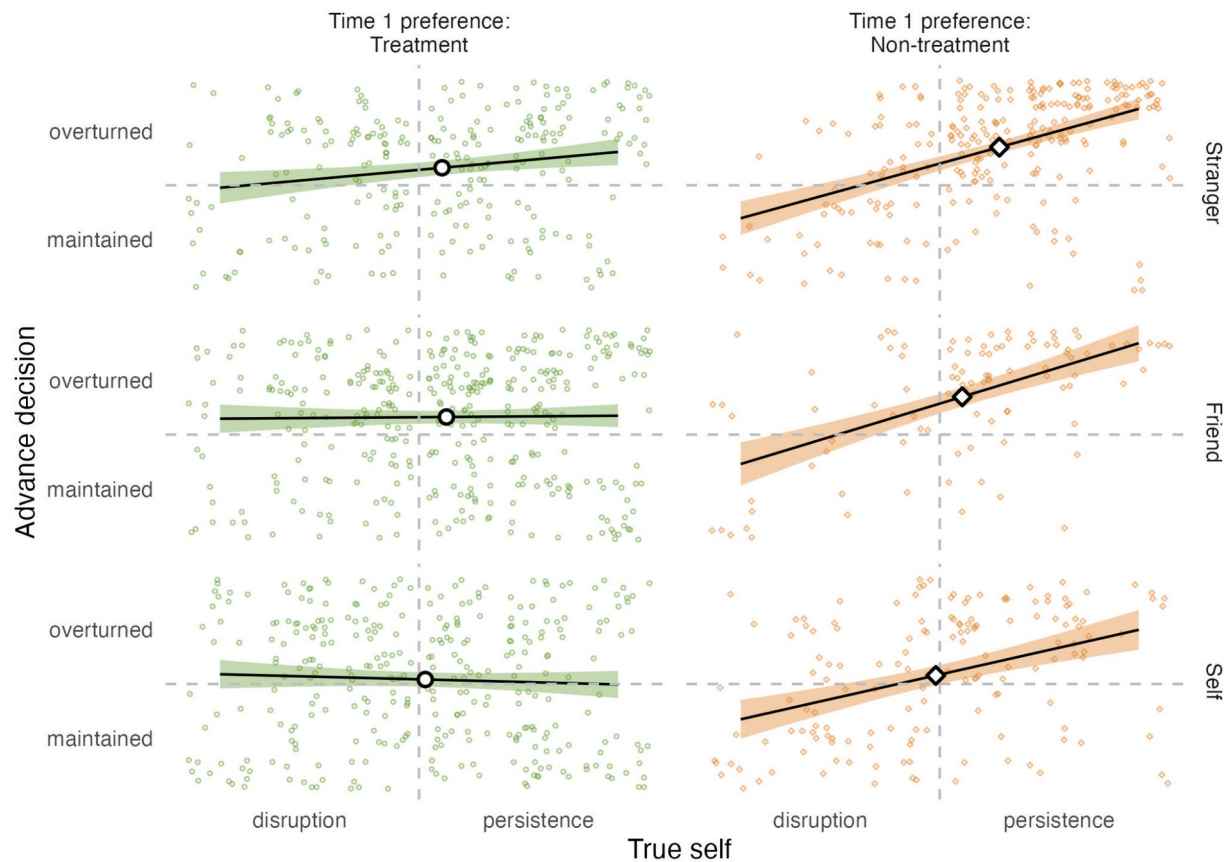


Figure 2. *Treatment Decisions by True Self Judgments for Each Target (horizontally) and Time 1 Preference (vertically): Scatter Plot, Linear Trend Line and Group Means.*

Entering true self ascriptions into a model of treatment decisions substantially improved fit to the data, $F_{(df = 6)} = 19.16$, $p < .001$ —revealing a main effect of true self ascriptions, $F_{(1)} = 21.16$, $p < .001$, $\eta^2 = .025$, a two-way interaction with Preference, $F_{(1)} = 85.55$, $p < .001$, $\eta^2 = .068$, and a small three-way Target×Preference×TrueSelf interaction, $F_{(2)} = 3.29$, $p = .038$, $\eta^2 = .005$. The interaction effects indicated that the relationship between true self ascriptions and treatment decisions varied across preference conditions and targets. When the T1 preference was to withhold treatment, true self ascriptions were associated with decisions to treat at T2 ($B = -0.56$, 95% CI [-0.67, -0.45]), and this effect was equally strong across conditions (pairwise $ps > .40$). Meanwhile, when the T1 preference was to receive treatment, true self ascriptions were associated with the decision not to treat the Stranger at T2 ($B = 0.21$, 95% CI [0.07, 0.36]), $t = 2.96$, $p = .003$, but were unrelated to treatment decisions in either the Best Friend or Self conditions, $ps > .19$.

3. Discussion

Consistent with our pre-registered hypothesis, we found that participants were more likely to defer to others' incapacitated T2 preferences than to contrary, capacitated T1 preferences, and that they were more likely to defer to the T2 preferences of others than themselves. However, unlike our hypothesis, the discrepancy between first- and third-person judgments did not depend on how the preference was phrased, suggesting that the hypothesized distinction by ambiguity was an artifact of the small sample size in our second exploratory pilot. The finding concerning ordinary people's tendency to defer to T2 preferences in third-person cases may have implications for surrogate medical decision-making.

First, the law recognizes many situations in which someone capable of expressing preferences may not have those preferences followed. For example, guardianship law considers whether the person generally lacks the capacity to understand the nature and consequences of their decision-making (Kohn 2020). Relatedly, according to standard bioethical accounts, capacity to make a treatment decision requires the cognitive ability to comprehend, understand, use, and consider information relevant to that decision (Foddy and Savulescu 2006; Holroyd 2009; Schaefer et al. 2014; Lewis 2021). Thus, in the cases we examined, most traditional accounts would likely conclude that preferences made with full cognitive functioning at T1 ought to guide treatment decisions. In stark contrast, our study documented high levels of deference to T2 preferences across all cases, revealing a notable tension between lay moral intuitions about end-of-life care and traditional philosophical models.

Second, under the doctrine of substituted judgment, surrogate decision-makers are directed to make the decision the person would have made if capacitated (Jaworska 2017; Appel 2022b), and previous studies have shown mixed results as to how good surrogate decision-makers are at predicting these decisions (Shalowitz, Garrett-Mayer, and Wendler 2006; Sulmasy, Haller and Terry 1994). Our present study builds on this literature, showing that surrogate decision-makers might *systematically* err when tasked with determining what the patient, in full possession of their faculties, would have wanted. In particular, surrogate decision-makers might underestimate the circumstances in which the person would have wanted their T1 preference followed.

Moreover, the results reveal a sharp difference between deciding for oneself and deciding for others close to oneself—but limited differences, if any, between deciding for

socially close versus distant others. In most analyses, although Best Friend cases fell between Self and Stranger cases, they looked much more like Stranger cases than Self cases. It appears, therefore, that it is primarily the qualitative distinction between thinking about oneself and others that drives the main effect, rather than the quantitative social distance between the decision-maker and the future character. This pattern suggests that the effect likely arises from the radically different kind of epistemic access people have to themselves and the reasons for their own preferences as compared to others, and only secondarily from how much personal knowledge they have of others.

Third, in finding a connection between true self judgments and treatment decisions, this study adds to the growing body of evidence that ordinary people do not see entitlement to have medical decisions respected to be merely a function of cognitive capacity (Earp et al., 2023; Toomey 2021). Rather, it supports the finding that the continuity of the true self plays a significant role. According to traditional views, respect for decisions depends on the decision-maker's cognitive capacity. In this study, most participants deferred to a decision expressed *after* substantial cognitive decline. Moreover, they deferred to different degrees across perspectives, notwithstanding identical cognitive functioning, correlated with true self ascriptions. Thus, it seems that most people think of the entitlement to have decisions respected as not only a function of a cognition, but also of the authenticity of a decision to the decision-maker's true self and the reasons they find compelling, to which people may have privileged epistemic access when thinking about themselves. In other words, participants may be more likely to think their future self with dementia is not their true self, based on its contrast with their direct, present, perceived knowledge of their own true self, whereas, in third-

person cases, participants have less evidence to judge the nature of another person's true self at T1 and may instead presume the continuity of the true self at T2.

Correspondingly, it might be that participants who think that the T2 character is still their true self may believe they retain privileged authority to change their mind.

We also note that our study drew on a convenience sample of US Americans—imposing limitations on the generalizability of our findings. Future work should ascertain whether our findings emerge in nationally representative samples, in other languages, and across cultures, and whether they extend to questions arising outside the healthcare domain. In addition, to insulate participants' moral reasoning from concerns about legal obligation, we did not describe T1 preferences as having been memorialized in writing. In practice, however, whether advance medical decisions are contained in a legally valid document could have an effect on the comparative (moral) weight that people assign to advance and current/future patient preferences. Finally, the study involved only one medical decision, in which the stakes were largely stipulated, relatively straightforward, and involved life or death. The findings may not generalize to cases involving greater uncertainty and more complex questions of probability and value.

In short, this study found an important discrepancy between people's medical decision-making from first- versus third-person perspectives—when placed in a surrogate decision-making role, people tended to disregard patients' prior, capacitated preference in favor of their incapacitated, contemporaneous treatment preference as to whether to treat a fatal illness. Meanwhile, when placed in the patient's perspective, participants were comparatively more likely to wish their *own* capacitated preference be

upheld. This finding challenges dominant assumptions in law, healthcare, and bioethics, and reveals an unexplored challenge for the practice of surrogate decision-making.

Declaration of Interest Statement:

ACKNOWLEDGMENTS: The authors would like to thank the members of the Oxford BioXPhi Lab, a workshop audience at the Yale-Oxford-Jagiellonian BioXPhi Conference in Kraków, and three anonymous peer reviewers, whose comments and feedback immeasurably improved this paper.

AUTHOR CONTRIBUTIONS: All listed authors helped conceive and implement this study, and draft this paper.

FUNDING: This work was supported by the Uehiro Centre for Practical Ethics at Oxford University, and the Spanish Ministry of Science and Innovation (grant number PID2020.119791RA.I00).

CONFLICTS OF INTEREST: The authors report no conflicts of interest.

ETHICS APPROVAL: This study was approved by University of Oxford CUREC, reference R80692/RE001.

References

Appel, J.M. 2022a. A values-based approach to capacity assessment. *Journal of Legal Medicine* 42(1-2): 53-65.

Appel, J.M. 2022b. Substituted judgment for the never-capacitated: Crossing *Storar's* bridge too far. *Bioethics* 34(2): 225-231.

Aron, A., Aron, E.N., & Smollan, D. 1992. Inclusion of the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology* 63(4): 596-612.

Batteaux, E., Ferguson, E., Tunney, R.J. 2020. A mixed methods investigation of end-of-life surrogate decisions among older adults. *BMC Palliative Care* 19 (44).

Batteaux, E., Ferguson, E., Tunney, R.J. 2019. Do our risk preferences changes when we make decisions for others? A meta-analysis of self-other differences in decisions involving risk. *PLOS One* 14(5): e0216566.

Brock, D.W. 1995. Death and dying: Euthanasia and sustaining life: Ethical issues. In *Encyclopedia of bioethics* (vol. 1), ed. W. Reich, 563-572. New York: Simon and Schuster.

Buchanan, A. 1988. Advance directives and the personal identity problem. *Philosophy & Public Affairs* 17(4): 277–302.

Dresser, R. 1986. Life, death, and incompetent patients: Conceptual infirmities and hidden values in the law. *Arizona Law Review* 28(3): 373-405.

Dworkin, R. 1993. *Life's dominion. An argument about abortion and euthanasia*. London: Harper Collins.

Earp, B.D., Demaree-Cotton, J., Dunn, M., Dranseika, V., Everett, J.A.C., Feltz, A., Geller, G., et al. 2020. Experimental philosophical bioethics. *AJOB Empirical Bioethics* 11(1): 30–33.

Earp, B.D., Latham, S.R., & Tobia, K.P. 2020. Personal transformation and advance directives: An experimental bioethics approach. *The American Journal of Bioethics* 20(8): 72–75.

Earp, B.D., Lewis, J., Dranseika, V., & Hannikainen, I.R. 2021. Experimental philosophical bioethics and normative inference. *Theoretical Medicine and Bioethics* 42(3–4): 91–111.

Earp, B.D., Hannikainen, I.R., Dale, S., & Latham, S.R. 2023. Experimental philosophical bioethics, advance directives, and the true self in dementia. In *Experimental philosophy of medicine*, eds. A. De Block and K. Hens. London: Bloomsbury.

Foddy, B., & Savulescu, J. 2006. Addiction and autonomy: Can addicted people consent to the prescription of their drug of addiction? *Bioethics* 20(1): 1–15.

Frisch, L., Kneer, M., Krueger, J., & Ullrich, J. 2021. The effect of outcome severity on moral judgment and interpersonal goals of perpetrators, victims, and bystanders. *European Journal of Social Psychology* 51(7): 1158-1171.

Holroyd, J. 2009. Relational autonomy and paternalistic interventions. *Res Publica* 15: 321–336.

Horvath, J. & Wiegmann, A. 2022. Intuitive expertise in moral judgments. *Australasian Journal of Philosophy* 100(2): 342-359.

Jaworska, A. 1999. Respecting the margins of agency: Alzheimer's patients and the capacity to value. *Philosophy & Public Affairs* 28(2): 105-138.

Jaworska, A. 2017. Advance directives and substitute decision-making. *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta,
<https://plato.stanford.edu/archives/sum2017/entries/advance-directives>

- Kahane, G., Everett, J.A.C., Earp, B.D., Caviola, L., Faber, N.S., Crockett, M.J., & Savulescu, J. 2017. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review* 125(2): 131-164.
- Kneer, M. 2018. Perspective and epistemic state ascriptions. *Review of Philosophy and Psychology* 9(1): 313–341.
- Kohn, N. 2020. *Elder law: practice, policy, and problems* (2nd ed). New York: Aspen.
- Lewis, J. 2020. From x-phi to bioxphi: Lessons in conceptual analysis 2.0. *AJOB Empirical Bioethics* 11(1): 34-36.
- Lewis, J. 2021. Autonomy and the limits of cognitive enhancement. *Bioethics* 35(1): 15–22.
- Lewis, J., & Holm, S. 2022. Organoid biobanking, autonomy and the limits of consent. *Bioethics* 36(7): 742–756.
- Lewis, J., & Holm, S. 2023. Towards a concept of embodied autonomy: In what ways can a patient's body contribute to the autonomy of medical decisions? *Medicine, Health Care and Philosophy* 26(3): 451-463.
- Lewis, J., Demaree-Cotton, J., & Earp, B.D. 2023. Bioethics: Experimental approaches. In *Encyclopedia of the Philosophy of Law and Social Philosophy*, eds. M. Sellers and S. Kirste, 279–286. Dordrecht: Springer.
- McMahan, J. 2003. *The ethics of killing: Problems at the margins of life*. Oxford: Oxford University Press.
- Mirza, O.F. & Appel, J.M. 2024. Capacity reconceptualized: From assessment tool to clinical intervention. *Cambridge Quarterly of Healthcare Ethics* 33(1): 35-39.
- Newman, G., Knobe, J. & Bloom, P. 2014. Value judgments and the true self. *Personality and Social Psychology Bulletin* 40(2): 203-206.
- Persad, G. 2019. Authority without identity: Defending advance directives via posthumous rights over one's body. *Journal of Medical Ethics* 45: 249-256.
- Schaefer, G. O., Kahane, G., & Savulescu, J. 2014. Autonomy and enhancement. *Neuroethics* 7(2): 123–136.
- Shalowitz, D.I., Garrett-Mayer, E., & Windler, D. 2006. The accuracy of surrogate decision makers: A systematic review. *Archives of Internal Medicine* 166(5): 493-497.

- Shiffrin, S.V. 2004. Advance directives, beneficence, and the permanently demented. In *Dworkin and His Critics with Replies by Dworkin*, ed. J. Burley, 195-217. Oxford: Blackwell.
- Strohming, N., Knobe, J., & Newman, G. 2017. The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science* 12(4): 551–560.
- Sulmasy, D.P., Haller, K., & Terry, P.B. 1994 More talk, less paper: Predicting the accuracy of substituted judgments. *The American Journal of Medicine* 98(5): 432-438.
- Tobia, K.P., Buckwalter, W., & Stich, S. 2013. Moral intuitions: Are philosophers experts? *Philosophical Psychology* 26(5): 629-638.
- Tobia, K.P., Chapman, G.B., & Stich, S. 2013. Cleanliness is next to morality, even for philosophers. *Journal of Consciousness Studies* 20(11-12): 195-204.
- Toomey, J. 2022. Narrative Capacity. *North Carolina Law Review* 100(4): 1073-1134.
- Toomey, J. 2021. Understanding the perspectives of seniors on dementia and decision-making. *AJOB Empirical Bioethics* 12(2): 101-112.
- Tsoy, E., Possin, K.L., Thompson, N., Patel, K., Garrigues, S.K., Maravilla, I., Erhoff, S.J., & Ritchie, C.S. 2020. Self-administered cognitive testing by older adults at-risk for cognitive decline. *Journal of Prevention of Alzheimer's Disease* 7(4): 283-287.
- Vazire, S. 2010. Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality & Social Psychology* 98(2): 281-300.
- Walsh, E. 2020. Cognitive transformation, dementia, and the moral weight of advance directives. *The American Journal of Bioethics* 20(8): 54–64.
- Wrigley, A. 2007. Personal identity and advance statements. *Journal of Applied Philosophy* 24: 381-396.
- Zikmund-Fisher, B.J., Lacey, H.P., Fagerlin, A. 2008. The potential impact of decision role and patient age on end-of-life treatment decision making. *Journal of Medical Ethics* 34(5): 327-331.
- Zikmund-Fisher, B.J., Sarr, B., Fagerlin, A., & Ubel, P.A.. 2006. A Matter of Perspective: Choosing for Others Differs from Choosing for Yourself in Making Treatment Decisions. *Journal of Internal Medicine* 21(6): 618-622.

Table 1. Percentages of dichotomized treatment decisions and true self judgments by perspective and T1 preference.

		T2 Treatment Decisions		T2 True Self Judgments	
		T2-Treat	T2-Withhold	True self persists	True self disrupted
Stranger	T1-Treat	36%	64%	59%	41%
	T1-Withhold	78%	22%	75%	25%
Friend	T1-Treat	37%	63%	62%	38%
	T1-Withhold	77%	23%	70%	30%
Self	T1-Treat	46%	54%	53%	47%
	T1-Withhold	59%	41%	51%	49%

Table 1. Three-Way ANOVA of Treatment Decisions and True Self Judgments.

	Treatment Decisions				True Self Judgments			
	<i>F</i>	<i>df</i>	<i>p</i>	η^2	<i>F</i>	<i>df</i>	<i>p</i>	η^2
<i>Target</i>	0.50	2	.60	.012	11.17	2	< .001	.017
<i>Preference</i>	147	1	< .001	.093	2.36	1	.12	.002
<i>Ambiguity</i>	0.01	1	.94	.000	0.01	1	.94	.000
<i>Target×Preference</i>	18.34	2	< .001	.026	5.78	2	.003	.008
<i>Target×Ambiguity</i>	3.39	2	.034	.005	1.17	2	.31	.001
<i>Preference×Ambiguity</i>	10.02	1	.002	.007	3.27	1	.071	.002
<i>Target×Preference×Ambiguity</i>	0.34	2	.71	.000	0.92	2	.40	.001

Figure 1. *Frequency of Treatment Decisions (Left) and True Self Judgments (Right) for Each Target (horizontally) and Time 1 Preference (color-coded): Grouped Histograms, Means and 95% Confidence Bands.*

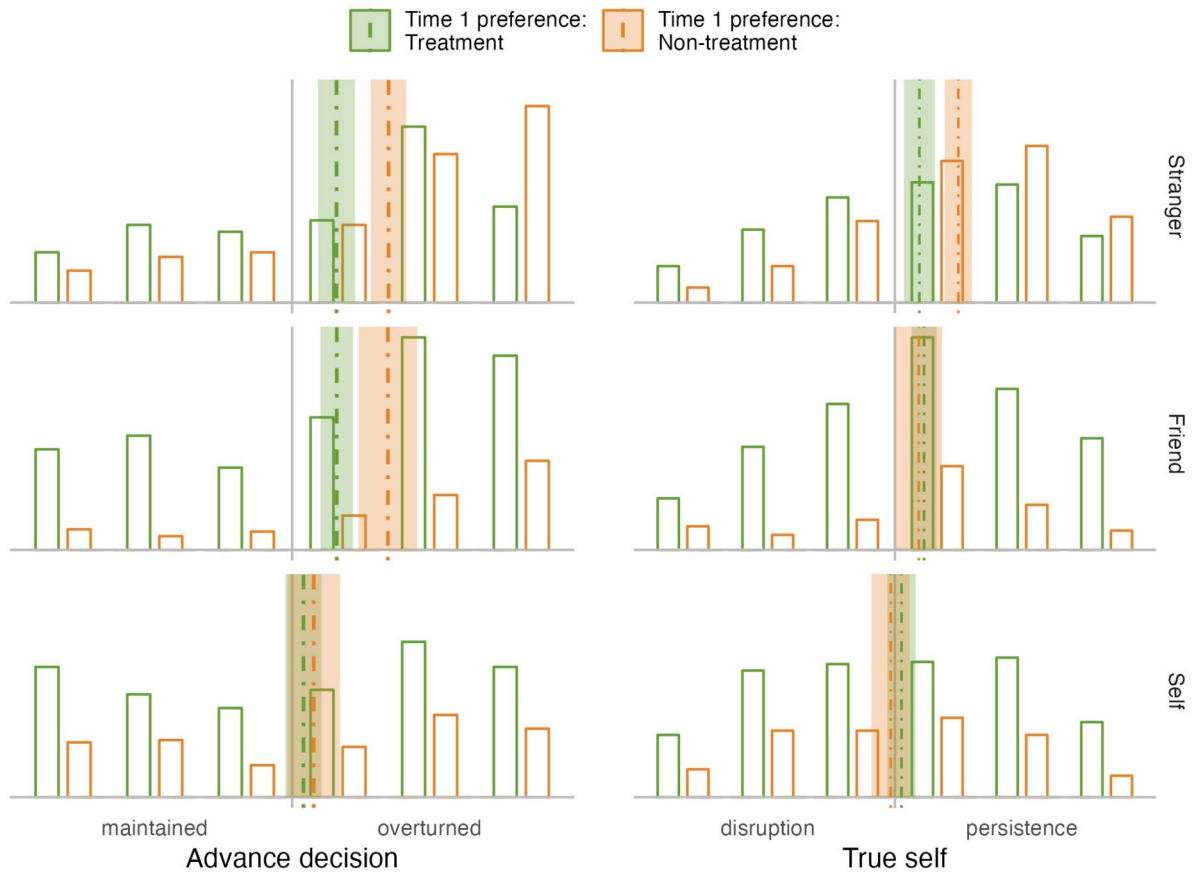


Figure 2. *Treatment Decisions by True Self Judgments for Each Target (horizontally) and Time 1 Preference (vertically): Scatter Plot, Linear Trend Line and Group Means.*

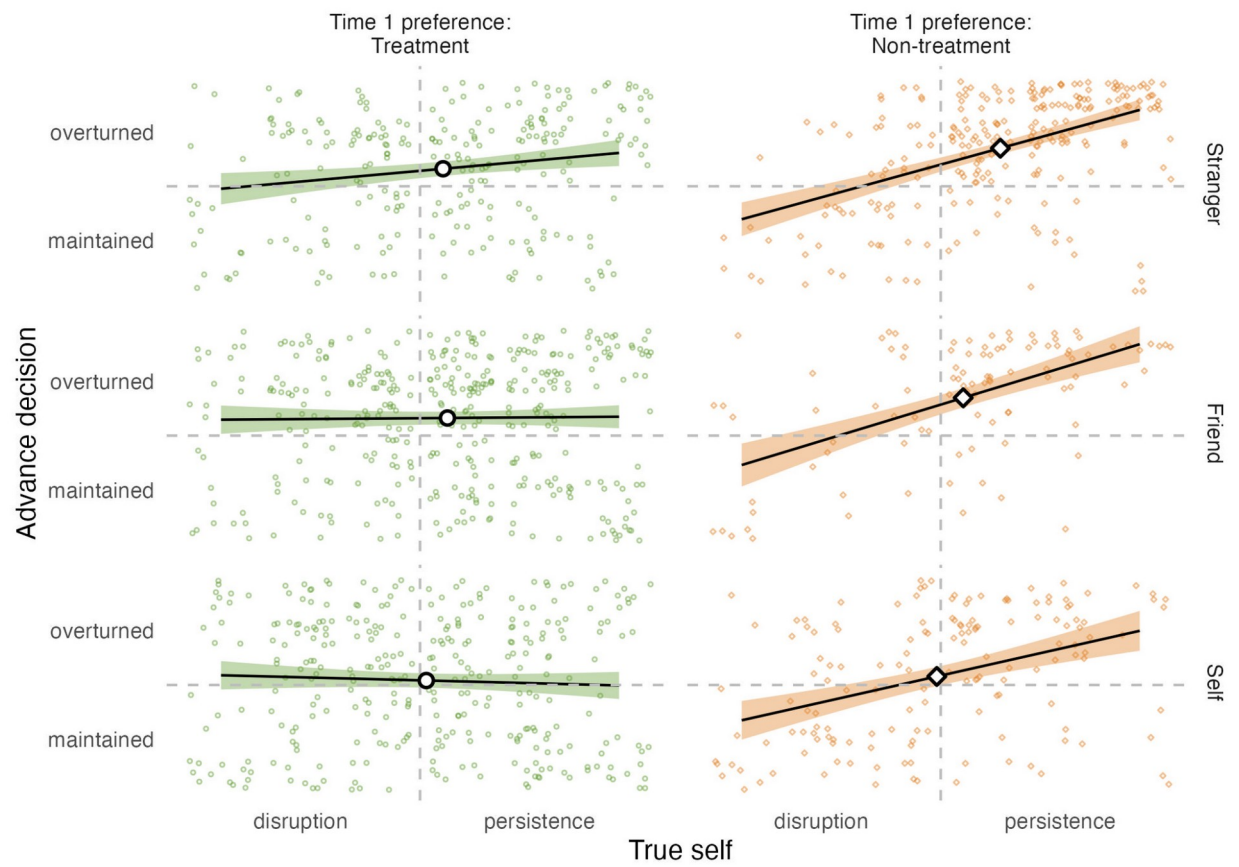


Figure Captions:

Figure 1. *Frequency of Treatment Decisions (Left) and True Self Judgments (Right) for Each Target (horizontally) and Time 1 Preference (color-coded): Grouped Histograms, Means and 95% Confidence Bands.*

Figure 2. *Treatment Decisions by True Self Judgments for Each Target (horizontally) and Time 1 Preference (vertically): Scatter Plot, Linear Trend Line and Group Means.*

Supplementary Analysis 1

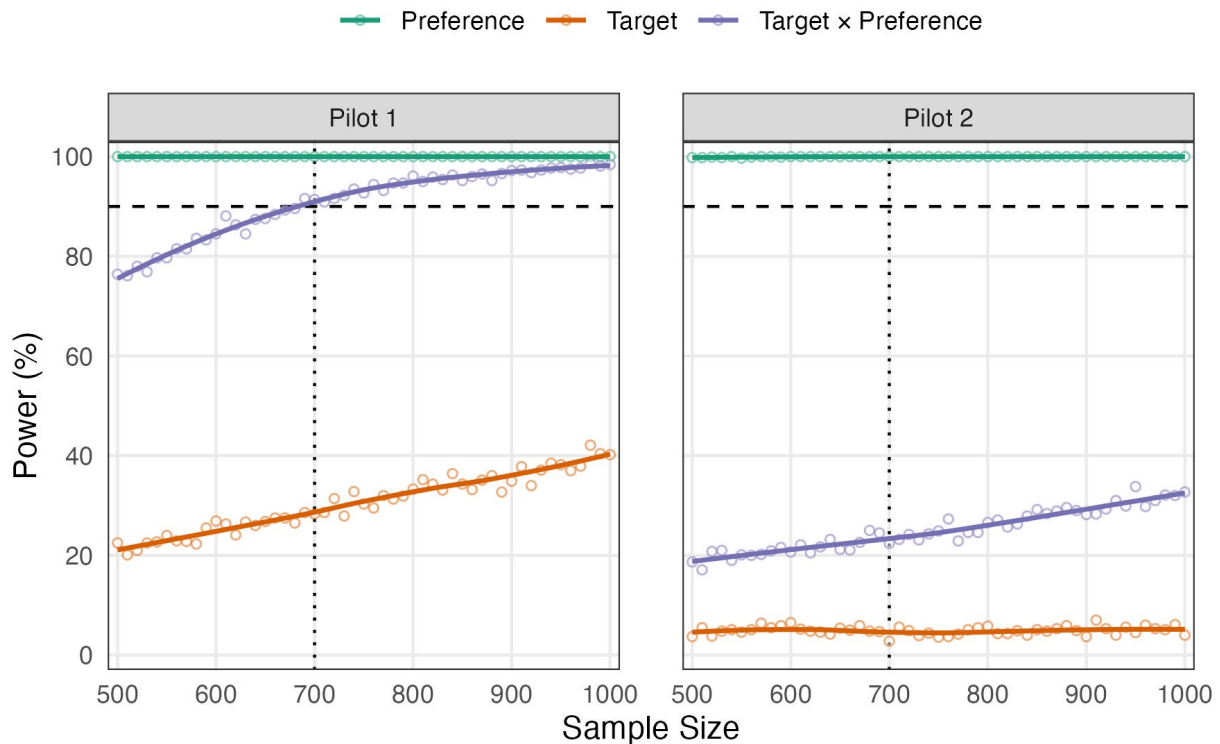
For each of two pilot studies ($ns = 202, 201$), we ran a two-way ANOVA of treatment decisions, with Target (Self, Friend, Stranger), Preference (Treatment, Non-Treatment) and the Target×Preference interaction. Both pilots revealed an effect of Preference (Pilot 1: $F_{(1, 196)} = 19.27, p < .001$; Pilot 2: $F_{(1, 195)} = 10.07, p = .001$), but no effects of Target or of the two-way interaction.

In the first pilot study, examination of the simple effects of Preference in each Target condition suggested that effect could be greater for the Stranger ($B = 1.35$), and especially a Best Friend ($B = 2.07$), than for the Self ($B = 0.77$)—despite the absence of a two-way interaction, $F_{(2, 196)} = 1.61, p = .20, \eta^2 = .016$. This effect then became the target of our power analysis (see Supplementary Analysis 2). When the T2 preference was phrased more ambiguously in the second pilot study, the magnitude of the difference across Preference conditions did not differ as clearly ($0.64 < Bs < 1.10$), $F_{(2, 196)} = 0.29, p = .75, \eta^2 = .003$.

Supplementary Analysis 2

To estimate the target sample size, we conducted a bootstrap power analysis by resampling from the pilot data with replacement (see <https://osf.io/snyuf>).

We established a range of simulated sample sizes from 500 to 1000 participants in ten-participant increments. For each sample size, n_i , we ran 1000 iterations of the target analysis (i.e., a two-way ANOVA), resampling n_i observations from the Pilot 1 dataset, and another n_i observations from the Pilot 2 data. In each iteration, we stored the p -values for all three terms in the ANOVA (Target, Preference, and Target $\times$ Preference). Finally, for each sample size and term in the model, we calculated the percentage of iterations that produced a significant effect (see Supplementary Figure 1). This approach resulted in a target sample size of 1400 participants (i.e., 700 per Ambiguity condition), which we increased by 100 to account for attention check failures.



Supplementary Figure 1. *Estimated Statistical Power for each Sample Size: Pilots 1 and 2. For the two-way interaction in Pilot 1, the dashed horizontal line (at 90% power) intersects with the estimated power trend line as the sample size approaches 700 (see the dotted vertical line).*